Housing, Environment and Cardio-respiratory Health: The Relative Influence of the Past and the Present

Jeremy Walker

PhD

The University of Edinburgh

2009

DECLARATION

As required by the University of Edinburgh Regulations in force at the time of submission, I declare:-

- that this thesis has been composed by me
- that the work reported is my own
- that the work has not been submitted for any other degree or professional qualification

Jeremy Joseph Walker

ABSTRACT

The existence of socially-patterned health (with poorer health generally being experienced by those in more disadvantaged circumstances) is widely recognised. Social differentials have been observed for (*inter alia*) respiratory disorders, and for cardiovascular disease. One possible explanation for social inequality in these areas of health posits a mediating effect of housing conditions: disadvantaged individuals may face greater exposure to residential hazards (such as dampness), which may in turn adversely influence cardio-respiratory health. However, few studies have examined a complete posited causal chain linking socioeconomic position (SEP) with health via housing.

Using pre-existing data, this study constructed detailed representations of the social and residential experiences over adult life (15 to 60 years) of a sample of elderly British people. Both measures of accumulated exposure (to disadvantage, and to housing hazards), and explicit trajectories of social and residential experience, were derived. Construction of trajectories required the development of methods for condensing individuals' diverse experiences into higher-level groups, in the interests of analytical tractability. Relationships between the derived measures of lifetime exposure and a range of outcomes expressing aspects of cardio-respiratory health in old age were assessed. No persuasive evidence was observed to support the hypothesis that lifetime residential exposures may mediate the relationship between SEP and the health outcomes examined. In addition to testing this specific conceptual model, the study examined how exposure to social disadvantage and to residential risks varied over adult life, identifying distinctive features of the exposure experience which could not readily be captured by the infrequent sampling of SEP commonly featured in health inequality research. The respective merits of such 'sparse' sampling and the more intensive sampling used in the study were compared. It was concluded that fully exploiting the additional information captured by intensive sampling

requires confronting a number of methodological challenges. Because of this, it is argued that the collection of detailed information on exposures over time does not automatically confer genuine advantages over the hitherto dominant approach of sampling at only a small number of time points. Future development of lifecourse epidemiology will require further debate over how lifetime exposure (to both social and environmental risk factors) can most effectively be represented in quantitative analysis.

ACKNOWLEDGEMENTS

Sincere thanks are due to my three supervisors for this study: Prof. Steve Platt (The University of Edinburgh), Prof. Richard Mitchell (The University of Glasgow) and Prof. David Blane (Imperial College, London). All gave unhesitatingly of their time and expertise, and the study would have been impossible without their very considerable contributions.

Thanks are also due to the Chief Scientist Office of The Scottish Government, who provided funding for the studentship under which the research reported here was carried out.

CONTENTS

LIST OF FIGURES, TABLES, BOXES AND APPENDICES

NOTE: Captions to figures etc. in the main text are sometimes lengthy; in the following list, the original captions are abbreviated in the interests of conciseness

PREFACE

The study reported here explored one possible route via which the persistently observed phenomenon of socially-patterned health might be explained. The underlying hypothesis tested was that relationships between socioeconomic position and health are mediated via the effect of housing conditions (such as exposure to residential dampness). The first four chapters of this thesis outline the background to the problem, beginning with a discussion of social inequality in health and culminating (in Chapter 4) in a formal statement of the research questions which were examined by the study. Chapter 5 describes the properties of a pre-existing dataset which was selected to facilitate investigation of these questions, while Chapters 6 to 8 describe how elements from this dataset were manipulated to provide the measures required by the study. The techniques adopted to assess the main associations of interest are described in Chapter 9, and a final methodological section (Chapter 10) outlines methods which were developed to explore subjects' experiences over time of the factors which were of main interest. Chapters 11 to 15 present results, with these being discussed in Chapters 16 to 18. The final chapter (Chapter 19) presents conclusions.

In contrast to some doctoral thesis submissions, there is no 'literature review' section *per se*; rather, relevant literature is introduced and discussed where appropriate. This occurs mainly in the four *Introduction* chapters, but further discussion of appropriate literature takes place at other points (for example, a substantial number of methodology-related references are introduced in the *Methods* chapters). A note on the general approach adopted to locate relevant literature appears towards the end of the thesis, immediately prior to the *References*.

Material from this study was presented at the 52nd Annual Scientific Meeting of The Society for Social Medicine (17-19 September, 2008). The relevant Abstract appears as Appendix 8.

CHAPTER 1: INTRODUCTION (I) - SOCIAL INEQUALITY IN HEALTH

1.1 Socioeconomic position and health: a ubiquitous relationship

Empirical evidence of social differences in health (generally in the form of poorer health being experienced by those in less favourable circumstances) has accumulated over an extended period of time. Evidence demonstrating strong associations between socioeconomic position ('SEP') and health dates as far back as ancient Greece, Egypt and China (Krieger *et al.*, 1997), and in mediaeval Europe exceptionally high rates of disease among miners were identified by Paracelsus (Lynch, 1996). More recent historical developments have been considered by Drever & Whitehead, who conclude that "relatively firm evidence of substantial social inequalities in mortality goes back to the 17th century in Geneva, and to the 18th century in other parts of Europe and Britain." (Drever & Whitehead 1997, p. 12). The evidence base for the existence of links between socioeconomic position and health continued to grow throughout the 19th and 20th centuries. Examples highlighted by Lynch (1996) include investigations into the relationship between rent levels and mortality in Paris in the 1820s, and Virchow's reporting of associations between poor living conditions and typhus in Upper Silesia in the 1840s. In the UK, an enquiry by the Poor Law Commission reported that in Liverpool in 1840 the mean age at death was 35 years for the upper and professional classes, as against 22 for tradesmen and 15 for manual workers and servants (Macintyre, 1997). Further historical developments in the accumulation of evidence for what came to be known as health inequality are identified by Drever & Whitehead (1997), Macintyre (1997) and Davey Smith *et al*. (2001).

A prominent modern landmark in the field of research into health inequality was the publication in the UK in 1980 of the Black Report (Department of Health and Social Security, 1980). This placed health inequality "firmly on the map of both public policy and academic study" (Bartley 2004, p. 1). The highly influential nature of the Report has been recognised by other commentators. For example, Lynch states that "publication of the Black Report in England in 1980 acted as a rallying point to reinvigorate [the] long research tradition of investigations into social conditions and health." (Lynch 1996, p. 21). Macintyre confirms the major impact of the Report on research into social inequalities in health, and defines its place in the historical landscape:-

> "It was preceded by some 140 years of concern with socio-economic differentials in death rates, and it was followed by an intense period of empirical and conceptual research on the subject." (Macintyre 1997, p. 723)

A major contribution of the Black Report was that it offered one of the first truly systematic presentations of a framework of possible explanations for the persistently-observed social patterning of health. This is not considered further at present, but forms the starting point for the discussion of competing explanations for social differentials in health which is presented in Chapter 2 below.

The phenomenon of socially-patterned health is currently regarded, by both policymakers and researchers, as almost ubiquitous:-

> "The invariable pattern, across time and between societies, is one in which men and women in higher socio-economic groups enjoy better health across longer lives than those in lower socio-economic groups." (Graham 2002, p. 240)

In fact, while this statement is broadly accurate, it represents a slight simplification. Historical examples of 'inverse' differentials in the social patterning of health (that is, circumstances in which the more affluent or those holding higher social status were exposed to greater risk of specific diseases or causes of death) exist. Bartley (2004) cites the elevated prevalence of coronary disease and angina which was observed among higher occupational classes in the UK during the period 1949-1953, while Macleod *et al*. (2005) highlight the increased risk of death in aircraft accidents experienced by more affluent people during the latter part of the 20th century. However, such instances of inverted health inequality are relatively rare, and Graham's statement reflects an almost universally accepted view. The very substantial evidence base for the existence of health inequality as it is conventionally understood (i.e. lower SEP co-existing with poorer health, whether assessed via mortality or morbidity) has been summarised in a number of reviews, including those by Macintyre (1986) and Davey Smith *et al*. (1990; 1994). The general acceptance of health inequality is so pervasive that reports of research in the field frequently begin with an almost cursory assertion that the phenomenon exists, as the following opening sentences from the published accounts of four studies show:-

> "The relation between low income and poor health is well established." (Lynch *et al.* 1997, p. 1889)

> "Inequalities in health associated with socioeconomic status are large and they are growing." (Kennedy *et al.* 1998, p. 917)

> "The association between adult social class and mortality is well established." (Pensola & Martikainen 2003, p. 745)

> "Socioeconomic inequalities in health and illness have been widely documented." (Laaksonen *et al*. 2007, p. 776)

A notable feature of health inequality is that it generally manifests itself as a gradient rather than a simple contrast between a disadvantaged group (who experience poor health) and the remainder of the population:-

> "…in country after country, study after study, what we see is not a group of very poor people at the bottom of the income distribution who have poor health while everyone else is fine. Instead, what we see is a steady gradation from the very top to the very bottom… …However many fine gradations of socioeconomic advantage anyone has been able to measure, these have so far all tended to show similar gradation in health". (Bartley 2004, p. 79)

The graded nature of health inequality has been recognised by other commentators including Marmot *et al*. (1997a) and Davey Smith *et al.* (1994), the latter referring to the 'fine grain' of socioeconomic differentials in health. An example of a gradient effect in health inequality is provided by mortality from diseases of the respiratory system among men in England and Wales between 1991 and 1993 (see Table 1.1). It is evident that, with a few deviations, death rates in this population tended to increase monotonically with lower SEP.

TABLE 1.1: Standardised mortality ratios (X 100) for deaths from diseases of the respiratory system (ICD Ninth Revision, Chapter VIII) for men in England and Wales, 1991 – 1993. Reproduced from Drever & Whitehead 1997, p. 134 (Table 10.8).

social class	**all diseases**	**pneumonia**	**chronic airway obstruction**	**bronchitis and emphysema**	**asthma**
I	42	58	21	44	51
II	56	69	42	43	55
IIINM	92	106	78	81	90
IIIM	115	93	131	125	128
IV	128	108	146	137	114
V	248	197	298	268	229

General acknowledgement of the existence of social inequality in health is accompanied by a widely (but not universally) held assumption that the relationship between SEP and health is a directional or causal one; that SEP *influences* health. A typical expression of this view is given by Williams: "We have clear, abundant evidence for a strong causal relationship between socioeconomic position and health." (Williams 1990, pp. 94-95). No further comment on this is made at present; possible explanations for the phenomenon of socially-

patterned health (including 'selection' theories which posit a reverse causation effect i.e. that health actually determines SEP) are considered in Chapter 2.

Irrespective of whether the relation between SEP and health status is regarded as causal, the concept of socioeconomic position which is central to that relation has found expression in many different forms in health research. The next section considers some of the approaches via which the abstract notion of SEP has been measured and represented in studies exploring the phenomenon of health inequality. However, before proceeding it is appropriate to conclude this introductory section by acknowledging that although this thesis is concerned solely with *social* differentials in health, inequality in health is evident along dimensions other than that of SEP. For example, patterning of health by ethnicity and by gender is discussed by Bartley (2004). Indeed, separating the respective effects of SEP and another potential determinant or predictor of health (such as ethnicity) can be challenging because the two factors may be inter-related:-

> "Ethnic minority groups, in general, do have a lower socio-economic status than the 'majority' population in the host country. Given the well-known association between socio-economic status and health, it is not surprising that ethnic inequalities in health are, to at least some extent, socio-economic in nature. Many empirical studies support this hypothesis." (Stronks & Kunst 2009, p. 1)

While inequality along dimensions other than SEP is not considered further in this thesis, it is helpful to bear in mind that confounding effects such as that identified by Stronks & Kunst are among the many factors which contribute to the difficulty of establishing the true causes of the persistently-observed relationship between SEP and health.

1.2 Measuring socioeconomic position

Despite its centrality to health inequality, a concise single definition of socioeconomic position is elusive: "Although researchers have an intuitive sense of what SEP means, the numerous ways of measurement indicate the complexity of the construct." (Galobardes *et al*. 2006a, p. 7). A variety of alternative expressions are used (often interchangeably) to express the concept, including social class, social status, and socioeconomic status. The rather loose use of such terms is specifically highlighted by Bartley (2004). This issue of terminology has been treated by Krieger and colleagues, who make a persuasive case for standardising on *socioeconomic position*: "We employ this term, rather than the more commonly used phrase 'socioeconomic status', because 'socioeconomic status' blurs distinctions between two

different aspects of socioeconomic position: (a) actual resources, and (b) status, meaning prestige- or rank-related characteristics." (Krieger *et al.* 1997, p. 346). The force of this argument is recognised by Galobardes and colleagues in a recent detailed review of indicators of SEP (Galobardes *et al*. 2006a; 2006b). The present thesis largely follows Krieger's suggestion of using socioeconomic position as the preferred designation[1]. However, when discussing the specific measure of SEP which is used in the analyses conducted for the study (namely the UK Registrar-General's occupational social class scheme; see Section 1.3), the alternative terms 'social location' and 'social position' are also used. The employment of these phrases is justified on the grounds that the occupational class scheme "is based on the notion of the general standing of an occupation within the community." (Davey Smith *et al*. 1992, p. 1556). This approach matches that of Bartley: "I also use the term 'social position' to refer to class and status." (Bartley 2004, p. 23)

The articles by Galobardes and colleagues cited above discuss the theoretical basis, measurement, interpretation and strengths / limitations of a number of widely-used indicators of SEP. Their account defines a broad typology of SEP indicators, under which measures are grouped into four main categories representing respectively education; income; housing tenure, housing conditions and household amenities; and occupation based measures[2]. These main groups are augmented by additional categories, including proxy indicators which may be used when direct measures of SEP are not available. One such proxy indicator is number of siblings, justified on the basis that "in some contemporary industrialised societies larger numbers of children are associated with poorer SEP." (Galobardes et al. 2006b, p. 98). A further class of indicator consists of ecological or area-level measures, which express the aggregated characteristics of geographical locales rather than of individuals. Examples of such ecological indicators are the deprivation indices devised by Jarman, by Carstairs and colleagues, and by Townsend and colleagues (Bartley & Blane, 1994). The four main groups of SEP indicators defined by Galobardes *et al*. are adopted to organise the material in the remainder of the present section, without further specific citation of this source. The objective of presenting this material is to demonstrate the wide variety of SEP measures which have been used in health inequality research, by providing a limited number of specific examples drawn from each group of indicators. No attempt at a systematic

[1] Exceptions to this rule are made when citing authors who use alternative terms such as 'socioeconomic status'. In such cases the cited authors' own terminology is retained.
[2] Some commentators argue that the most commonly-used indicators of SEP fall into only three main types: education, income and occupational status (e.g. Williams & Collins, 1995; Dalstra *et al.*, 2006). However, the prominent role played by housing conditions in the present study makes the four-way classification of Galobardes and colleagues more useful.

enumeration of all possible indicators is made; rather, the goal is to provide a flavour of the multiplex ways in which the abstract concept of SEP has been realised.

A first group of SEP indicators consists of measures of education, generally representing either the number of years' education which the subject (or her / his parent[s]) has received, or the attainment of specific educational milestones (such as the award of a university degree). One fairly complex example of an education-based indicator of SEP featured in a study by Bronnum-Hansen & Baadsgaard (2007). These authors combined data on subjects' schooling, vocational training and further education to construct a three-way ordinal scheme which classified the individual's educational level as one of low, medium or high. Differences in life expectancy across these groups were then investigated. A study of Glymour *et al*. (2008) used both parental education (dichotomized at 8+ years) and personal education (number of years) as adjusting factors representing socioeconomic status when investigating the effect of race on stroke risk. Ramsay and colleagues (2008) used age at leaving full-time education as one measure of SEP in assessing the extent of social inequalities in disability among elderly British men. As a final example, a recent study by Loucks *et al*. (2009) used both parental education (expressed as a three-way ordinal scheme) and the subject's own length of education (number of years, collapsed into three ordered groups) when investigating associations between lifetime SEP and the incidence of coronary heart disease.

A second group of commonly-used indicators of SEP represents measures of income, generally expressed either as absolute income or as one of a range of predefined categories within an ordinal classification (e.g. '£5,000 to £9,999'). Frequently, the measurement captures the income of the household rather than that of the individual subject. One example of the use of income as an indicator of SEP in health inequality research is provided by a study of Lynch *et al*. (1997). In this, information on household income at three time points was used to determine the number of occasions (minimum zero, maximum three) on which participants were considered to experience economic hardship (defined as household income less than 200 percent of the US federal poverty level for the years involved). Associations between this quantity and measures of physical, cognitive, psychological and social functioning were assessed. A second illustration of the use of income as an indicator of SEP is given by a very large-scale study of Rogot and colleagues (1992), which used data on *c*. 800,000 white persons from the US National Longitudinal Mortality Study to assess relationships between family income (collapsed into seven bands) and life expectancy. As a

final example, an investigation by Meng and colleagues (2008) of the determinants of frequent asthma symptoms among low-income individuals featured a dichotomous indicator of family income, classifying subjects as below or above the 2001 US federal poverty level. This binary marker was used to determine the respective influences of traffic-related exposures, poverty and 'vulnerabilities' (such as health insurance coverage and employment status) on the experience of frequent asthma symptoms.

A concept related to income is that of 'wealth'[3], which arguably offers certain advantages over income as an indicator of individual economic status:-

> "While income is a current measure of a person's economic position, wealth is a much better indicator of their economic position in the long term, as wealth is accumulated throughout the life-course." (Shaw *et al*. 1999, p. 149)

The volatility of income (and hence its potential weakness as a measure of SEP) is highlighted by other commentators (e.g. Williams & Collins, 1995). An instance of the use of wealth in a health inequality context is provided by a study of Avendano *et al*. (2009) which compared the health of rich and poor Americans with that of Europeans. These authors constructed a complex measure of household 'wealth' which included financial assets such as stocks, bonds and savings, together with the value of concrete major owned items such as the participant's main residence and car(s). The measure was used to examine between-country health variations by wealth, and to investigate how associations between wealth and health differed across regions. One limitation of wealth as an indicator of SEP is that (unlike its inverse i.e. poverty) the concept of wealth and its measurement is relatively unresearched:-

> "While there is a considerable body of work on poverty, including a wide range of approaches that attempt to measure poverty, this is not matched by the amount of literature that addresses wealth. Indeed, it is remarkable how little literature there is on the subject." (Dorling *et al*. 2007, p. 3)

A third group of SEP indicators is based on housing tenure, housing conditions and household amenities. There is some degree of conceptual overlap between this group and the concept of wealth introduced above, in that housing tenure may be seen as a reflection of individual wealth (Shaw *et al*., 1999). This group includes a subset of indicators representing housing conditions (that is, physical characteristics of the residential

[3] The typology of Galobardes and colleagues actually treats measures of wealth as being distinct from indicators of income, while recognising that the latter is a component of the former.

environment such as the presence of dampness and condensation). This concept is of central importance to the present study, and is considered separately in detail in Chapter 3. An example of the use of housing tenure in health inequality research is provided by a study of Wannamethee & Shaper (1997) which investigated associations between SEP and mortality in middle-aged British men. In this study, subjects' housing tenure (specifically, home ownership) was combined with car ownership to create a composite measure of material assets which was then related to the experience of mortality. Dalstraa and colleagues (2006) used housing tenure (dichotomized as owners *vs.* renters) in a study aimed at determining the effectiveness of a number of socioeconomic indicators as predictors of less than good health among elderly people. Although tenure has been widely used as an indicator of SEP in health research, it suffers from difficulties of interpretation. Ellaway & Macintyre (1998) have suggested that while tenure has consistently exhibited associations with longevity and with a number of health measures, these effects may be partly attributable to the fact that different tenure categories expose people to different levels of residential hazards (such as dampness):-

> "...rather than simply acting as a marker of material well being, housing tenure may be telling us something about exposures to health promoting or health damaging features of the dwelling". (Ellaway & Macintyre 1998, p. 148)

This possibility (that housing quality may operate as a mediating influence in the relationship between SEP and health) is central to the hypothesis tested by this study (see Chapter 4).

The remaining set of indicators in this group - household amenities - represents the availability or ownership of features such as central heating and sole use of bathrooms. However, as the prevalence of these amenities grows in modern developed societies, their usefulness as socioeconomic indicators declines. To illustrate, knowing that an individual lives in a home with its own private bathroom may be of limited value if the same is true of almost everyone else. One amenity which has featured extensively in health research is ownership of (or access to) motor vehicles. An example is provided by a study of Harding *et al*. (1997) which found that (with certain limited exceptions) mortality among elderly men and women was uniformly higher among those without access to a car than among individuals with such access.

The fourth group of indicators in the typology of Galobardes and colleagues includes measures based on the individual's occupation. The sociological theories which underpin these occupation-based measures of SEP are discussed in detail by Bartley (2004), and are

not rehearsed here. Commonly-used occupational indicators include the Erikson and Goldthorpe class schema (Erikson & Goldthorpe, 1992), and three measures which were developed by UK government bodies (specifically the General Register Office and its successors, the Office of Population Censuses and Surveys and the Office of National Statistics). These three measures are (in order of their introduction) Social Class based on Occupation ('SC'); Socio-economic Groups ('SEG'); and National Statistics Socio-economic Classification ('NS-SEC'). A historical outline of these three indicators, and of their mutual relationships, is given by Rose *et al*. (2005). Of these, SEG has been used in many academic studies (Rose *et al*., 2005), while NS-SEC (which was developed from the Erikson and Goldthorpe schema [Office for National Statistics, 2005]) has been used in official UK statistics and surveys since 2001. However, SC is of particular interest within the present study, because the measures of social position used in the study are based on the SC classification. The SC scheme is "the most commonly used individual socioeconomic indicator in British studies" (Davey Smith *et al*. 1998, p. 400); its position as the most commonly used measure of social position in British health inequality research is confirmed by Bartley (2004). Due to its central role in the present study, a discussion of the SC scheme is undertaken separately (see Section 1.3 below). Three examples chosen from the very large number of studies which have used SC to assess relationships between SEP and health now follow. Ford and colleagues (1994) used the occupational class of subjects' heads of household in an investigation of social gradients in a range of directly measured and self-reported dimensions of health. Hart *et al.* (1998), in a study of mortality among Scottish men, captured the occupational class of subjects at three stages of life: childhood, entry to the regular labour market and the time of entry to the study. This longitudinal representation of SC[4] was then used to assess the contribution of SEP at different times of life to mortality from a range of causes. A methodological study conducted by Manor and colleagues (1997) included social class at birth (based on parental [father's] occupation) as one of two social indicators in a comparison of three different methods for measuring inequality.

***

The above discussion provides only a cursory illustration of the diversity of ways in which SEP has been operationalised in research. While superficial, the discussion serves to

[4] The analyses conducted by these authors actually used a condensed version of the SC measure, the original six groups defined in the scheme (see Section 1.3) being reduced to four. Such condensations are often employed when SC is used in health research; the reduction to a binary contrast between manual and non-manual status is particularly common.

illustrate the desired point: that the concept of SEP, which is central to the phenomenon of health inequality, is elusive, complex and rather difficult to define. The essentially multifactorial nature of SEP is recognised by commentators:-

> "...socio-economic position is a multidimensional concept. It includes key components such as educational level and occupational class, but also employment status, income level and other indicators for material welfare." (Stronks & Kunst 2009, p. 1)

> "The position that different groups occupy in the social structure can be based on occupation, income, sex, family relations, heredity, political or religious affiliation, caste, skin color, or the number of pigs that are owned. Simple measures such as income and education are imperfect markers of the economic processes that structure human life from before conception to death." (Lynch 1996, p. 21)

Because of its multi-faceted nature, socioeconomic position when featured as a factor in research presents obvious difficulties of measurement. It is clearly different from, say, a biomedical measure like blood pressure, or an anthropometric measure such as leg length. In the two latter cases, the quantity of interest could in theory be measured with absolute precision (though in practice limitations of the measuring device or its operator make full precision unattainable). However, with SEP there is a further layer of imprecision or 'error', in that it is not possible to determine with certainty how effectively a single chosen measure of the subject's socioeconomic position (such as number of years of education, or her / his current occupational class) represents that person's 'true' SEP (which is ultimately undefinable). For this reason, much research in the field of health inequality cannot aspire to the goal of what might be termed 'perfection in measurement', which is a laudable aspiration of research in the biological or physical sciences. Like most research which is based on assessing associations between SEP and some other factor(s), the present study is characterised by a degree of uncertainty over how far the representation of SEP used is actually meaningful. The point is a philosophical rather than a methodological one, and is not considered further in this thesis, but merits highlighting at this early stage.

As indicated earlier, the measures of SEP which were used in this study are based on the UK Registrar General's occupational class scheme (SC) which was introduced above. A fuller account of this widely-used indicator of SEP follows in the next section.

1.3 The UK Registrar General's occupational social class scheme

1.3.1 History

The history of the UK occupational class scheme is given in outline by Rose (1995); a detailed account (including a great deal of interesting background information) is provided by Szreter (1984). In essence, the introduction of the UK occupational classes in 1913 (in the 74th Annual Report of the Registrar General) was largely due to the work of one T H C Stevenson, a medical statistician employed at the General Register Office for England and Wales. Stevenson's original scheme classified individuals into one of eight categories. Of these, five were graded representations of occupation, of which three were explicitly defined. Class I consisted of the upper and middle classes; Class III held skilled workmen; and unskilled labourers were assigned to Class V. An intermediate class was interposed between Classes I and III, and a second between Classes III and V, but these two groups were not assigned descriptive labels. The remaining three groups in Stevenson's scheme were reserved for individuals employed in specific industrial sectors (mining, textiles and agriculture). In 1921, the three sector-specific categories were removed, leaving a five-way graded classification. The SC scheme underwent further changes thereafter, although these had little effect on the overall shape of the model. The final form of the SC scheme, and that which has been used extensively in health research, is as shown below (reproduced from Rose *et al*. 2005, p. 8 [Table 1]):-

class	I	-	Professional, etc, occupations
	II	-	Managerial and technical occupations
	III	-	Skilled occupations
		(N)	Non-manual
		(M)	Manual
	IV		Partly skilled occupations
	V		Unskilled occupations

As explained earlier, this representation of SEP is the most commonly used measure of social position in British health research. It has attained this dominant position despite being subject to a number of limitations which have been widely acknowledged, and are now summarised.

1.3.2 Limitations

One major limitation of SC is that it makes no provision for classifying those who do not have a formal occupation; that is, those who are not currently employed. Consequently, certain groups of individuals (including full-time students, retired persons and the unemployed) are effectively excluded from reliable classification under the scheme. Where SC is used in research, one approach to handling such groups is simply to eliminate them from analysis. However, this clearly limits the generalisability of research findings: if a (possibly substantial) proportion of the population of interest is disregarded, inferences drawn from the results of a study are potentially of limited applicability. A particular concern is that those not in employment will often include individuals with serious and persistent illness, who are prevented by that illness from holding down a regular job. Thus, health inequality studies which use SC as a measure of SEP will tend to exclude subsets of people (the chronically sick) who may actually be of the greatest interest. A second response to those who are not currently in employment is to assign the person's previous occupation (Galobardes *et al*., 2006a), though this introduces certain assumptions about the extent to which social circumstances persist across changes in occupational status. The treatment of non-employment in the present study is discussed in Chapter 6 in the *Methods* section of this thesis. One large and particularly problematical subgroup of those who are not in paid employment consists of women whose participation in the labour market is specifically limited by the demands of looking after children or running the household. Issues relating to this group are discussed further in Sections 6.1.1 and 6.1.3.

A second weakness of SC is that the assignment of individuals to classes on the basis of occupation is in some cases vulnerable to imprecision; for example:-

> "...someone known as an 'engineer' could be a shop-floor worker (in the skilled manual, social class IIIM group) or a professional engineer in social class I (generally a non-shop-floor worker with a degree)." (Shaw *et al.* 1999, p. 89)

Related to the above limitation is a third, applicable only to longitudinal studies which sample the subject's SEP at multiple time points. It has been argued that the range of occupations available to the workforce has changed appreciably from one decade to another, so that comparisons of SEP across decades on the basis of SC are difficult to interpret:-

> "Coding occupations to class II in the 1950s meant small tradesmen, shopkeepers, senior clerical officers, shop managers, a few teachers and academics. Today it is

> overwhelmingly an educated managerial technological class far removed from the corner shop and the stonemason's yard." (Illsley & Baker 1991, p. 360)

This characteristic of SC (that the occupation-to-class mappings have changed over time) is one potential contributor to the 'artefact' explanation of health inequality which was presented in the Black Report, and is considered along with alternatives in Chapter 2.

In addition to the three specific limitations outlined above (limited population coverage, imprecision in the assignment of occupations to classes, and shifts in the nature of these assignments over time), the SC scheme has been widely criticised on more general conceptual grounds. These criticisms have been articulated by Murgatroyd (1984) and Illsley & Baker (1991), and are not considered here. Despite its weaknesses, the occupational class scheme has been extensively used in British academic research. Supporting the previously-cited views of Davey Smith *et al*. and of Bartley (see Section 1.2), the widespread use of SC in research contexts is highlighted by Rose & O'Reilly: "The classifications [occupational class and socio-economic groups] are extensively used as off-the-peg tools in academic research." (Rose & O'Reilly 1998, p. 53). The status of SC in health research is further confirmed by Macintyre: "...the RG scheme has been remarkably persistent despite its (lack of) theoretical base." (Macintyre 1986, p. 400). The widespread use of SC in research is largely explained by the persistent finding that (crudely stated) 'it works'; that is, has been found consistently to predict a variety of health outcome measures:-

> "Although the subject of sustained criticism over the years, the RGSC has been shown to be a strong and consistent predictor of a range of life outcomes". (Sturgis & Sullivan 2008, p. 72)

> "…occupational class (whether of self, father or husband) has repeatedly been shown to be associated with a diverse collection of health measures, including death from all causes or from specific causes, physical and mental illness, height, weight for height, birth-weight, blood pressure, dental condition, ability to conceive and self-perceived health. These class variations in mortality, morbidity, and other health related variables have been reported extensively during the last decade in Britain". (Macintyre 1986, p. 395)

***

Discussion thus far has sought to emphasise three key points. The first of these is a near-truism: that social inequality in health exists, and is widespread. A second point is that socioeconomic position (which constitutes one 'half' of the health inequality construct) is by nature complex and not amenable to concise definition. This has been demonstrated by

illustrating the diversity of ways in which researchers have sought to measure the underlying concept. Finally, some acknowledged limitations of occupational class (the most commonly-used measure of SEP in British health studies) have been briefly introduced. The focus on this specific measure of SEP is justified on the grounds that it forms the basis of the analyses performed for the present study.

Attention now turns to explanations which have been proposed to account for the persistently-observed phenomenon of socially patterned health.

CHAPTER 2: INTRODUCTION (II) - EXPLANATIONS FOR SOCIAL INEQUALITY IN HEALTH

2.1 The explanatory framework of the Black Report: artefact and selection theories of health inequality

As indicated earlier, a notable feature of the Black Report was its detailed consideration of possible explanations for the existence of health inequality. Chapter 6 in the Report ('Towards an Explanation of Health Inequalities') advanced four possibilities, which were described respectively as 'artefact explanations', 'theories of natural or social selection', 'materialist or structuralist explanations' and 'cultural or behavioural explanations'[5]. The artefact explanations approach (which received little attention in the Report) was introduced thus:-

> "This approach suggests that both health and class are artificial variables thrown up by attempts to measure social phenomena and that the relationship between them may itself be an artefact of little causal significance." (Townsend *et al*. 1992, p. 105)[6]

In essence, this view argues that there is no real relationship between class and health, maintaining rather that "the way social class or health is measured might influence the apparent magnitude of, and trends in, observed inequalities in health" (Macintyre 1997, p.727). The Report also introduced a variation on the artefact theory, expressed thus:-

> "...the failure of health inequalities to diminish in recent decades is believed to be explained to a greater or lesser extent by the reduction in the proportion of the population in the poorest occupational classes. It is believed that the failure to reduce the gap *between* classes has been counterbalanced by the shrinkage in the relative size of the poorer classes themselves." (Townsend *et al*. 1992, p. 105)

Under this version of the artefact explanation, class differentials in health might have persisted but the increasing proportion of the population assigned to the higher classes (who generally enjoy better health) meant that the health of the population *as a whole* had actually improved. Thus, between-class differentials would provide an imperfect indication of time-related changes in overall population health. One possible contributor to this type of artefact

[5] A discussion by Macintyre (1997) of the Black Report's four explanatory models argues that each explanation may be subdivided into two different expressions (labelled 'hard' and 'soft' interpretations) of the underlying idea. There is little merit in transcribing or paraphrasing this lengthy discussion here, but Macintyre's treatment does identify a degree of imprecision in the Report's approach to explaining the phenomenon of health inequality.

[6] All quotations from the Black Report in this chapter are taken from the republished edition of 1992 (Townsend *et al*.).

effect is the change over time in the assignment of occupations to social classes which was highlighted in Section 1.3.2. A number of other possible types of artefactual effect have been identified by Davey Smith *et al.* (1994).

Research in recent years has largely moved on from the artefact theory, which on examination of the evidence has been judged unconvincing. A review by Davey Smith and colleagues considered the respective merits of the explanations advanced in the Black Report and concluded:-

> "The many possible forms of artefactual distortion of associations between socio-economic status and mortality risk are examined and judged to have little effect; if anything, artefactual factors mean that conventional ideas about the magnitude of socio-economic differentials in mortality are an underestimate." (Davey Smith *et al*. 1994, p. 131)

A second class of explanation advanced in the Black Report ('theories of natural or social selection') posited an effect of reverse causation, hypothesising that rather than low social position causing poor health, individuals in better health (or endowed with superiority in other attributes, such as intelligence) tend to move up the socioeconomic scale, while the unhealthy are subject to downward social mobility. Under this theory, the phenomenon of health inequality is in effect an expression of inequality in other factors, and occupational class "is seen as a filter or sorter of human beings." (Townsend *et al*. 1992, p. 105). As with the artefactual theory, explanations based on social selection have come to be largely discounted. The previously-cited review by Davey Smith and colleagues concluded that "extensive research has produced little evidence that either intragenerational social drift or direct intergenerational selection contribute in a major way to the explanation of health inequalities." (Davey Smith *et al*. 1994, p. 138). However, although selection currently receives less attention as an explanatory theory of inequalities, the wider concept of selection may still influence research in ways which are not immediately apparent. For example, the dataset used in this study (see Section 5.2) holds responses from *c.* 300 individuals who were aged between 63 and 78 years when health outcome information was elicited, and it may be argued that this sample exhibits *de facto* 'selection' on the grounds of superior health in that it consists of people who were still alive at these relatively advanced ages.

2.2 Current theories of health inequality

While the artefactual and selection theories offered in the Black Report are currently viewed as unconvincing, the Report's two remaining explanatory models persist as viable contenders to explain the social patterning of health. Of these, the materialist explanation (the theory given most credence in the Report) argues that differences in the material and physical conditions of life (such as housing conditions and exposure to occupational hazards) explain (or contribute to) class gradients in health. The final possibility identified in the Report ('cultural or behavioural explanations') posits that individuals in different social strata engage to varying degrees in behaviours (such as indulgence in, or abstention from, smoking or recreational exercise) which are detrimental to or protective of health, thereby accounting for social patterns in health. Since publication of the Black Report, these theories have been joined by two further main explanatory models which seek to account for social health differences. The first of these is the psychosocial model, which proposes that health differentials result largely from perceptions of relative disadvantage, and the psychological sequelae of such perceptions, rather than from the direct effects of varying material circumstances: "Economic and social circumstances affect health through the physiological effects of their emotional and social meanings." (Marmot & Wilkinson 2001, p. 1233). Finally, in a major conceptual development, the lifecourse approach (Kuh & Ben-Shlomo, 1997) asserts that health reflects "patterns of social, psychological and biological advantages and disadvantages experienced by the individual over time… …these patterns being profoundly affected by the position of individuals and families in social and economic structures." (Bartley 2004, p. 115). Currently, these four explanations (materialist, behavioural / cultural, psychosocial and lifecourse) are the main contenders hypothesised as accounting for the social differentials in health which are persistently observed (Bartley, 2004). Of these, the lifecourse approach is of central interest to the present study, which posits a specific conceptual model of the way in which social and environmental exposures acting across adult life may influence aspects of health in old age. Lifecourse theories have received extensive attention in the research literature. In addition to the seminal collection of articles edited by Kuh & Ben-Shlomo (1997), general discussions of the lifecourse approach (as distinct from single studies which test hypotheses consistent with that approach) include those by Wadsworth (1997; 2007), Ben-Shlomo (2002), Davey Smith (2003), Kuh *et al.* (2003) and Lynch & Davey Smith (2005).

The four general classes of explanation listed above include variants of the basic concept embodied by each. For example, the original materialist theory of health inequality has been

augmented by ‘neo-materialist’ explanations which focus on “a combination of negative exposures and lack of resources held by individuals, along with systematic underinvestment across a wide range of human, physical, health, and social infrastructure.” (Lynch *et al*. 2000, p. 1202). Similarly, lifecourse explanations may broadly be divided into ‘critical period’ approaches (which assume that adverse experiences at particular ages have substantial health impacts later in life), and ‘accumulation of risk’ models, which hypothesise that progressive exposure to harmful events and circumstances across the lifecourse increases the risk of chronic disease and mortality (Ben-Shlomo & Kuh, 2002). These two main strands of the lifecourse paradigm may themselves be further refined. For example, Ben-Shlomo & Kuh define a typology of conceptual lifecourse models which takes the following form (reproduced from Ben-Shlomo & Kuh 2002, p. 287 [Table 1]):-

Critical period model
- with or without later life risk factors
- with later life effect modifiers

Accumulation of risk
- with independent and uncorrelated insults
- with correlated insults
  - ‘risk clustering’
  - ‘chains of risk’ with additive or trigger effects

The respective merits of the four general explanations for health inequality have been extensively debated, proponents of each presenting the case for her or his preferred concept. For example, Davey Smith and colleagues applied the Black Report’s original four conceptual models to mortality data and identified materialist factors as an attractive explanatory category, but argued the case for a lifecourse approach: “Progress in this area [socioeconomic differentials in mortality] will depend upon studies which can examine how exposures interact and accumulate over the course of life to produce the observed pattern of mortality risk.” (Davey Smith *et al*. 1994, p. 131). In a directly competitive approach, a pair of related papers by, respectively, Lynch and colleagues (Lynch *et al*., 2000) and Marmot and Wilkinson (Marmot & Wilkinson, 2001) debated the evidence inconsistent with and supportive of the psychosocial theory of inequality.

Although such debates continue, one important development has been a growing acceptance that rather than any single conceptual model providing a comprehensive, definitive general explanation for the social patterning of health, it is possible that each may embody a degree of truth. For example, Marmot and Wilkinson (advocates of the psychosocial theory of health inequality) concede the contribution of purely material factors to the maintenance of socially patterned health: “there are psychosocial pathways associated with relative

disadvantage which act in addition to the direct effects of absolute material living standards." (Marmot & Wilkinson 2001, p. 1233). In a similar vein, Bartley comments: "But these explanations [of health inequalities: material, cultural-behavioural, psychosocial and lifecourse] do not have to be mutually exclusive. It is likely that they need to be understood in combination with each other, though this creates a very complex task." (Bartley 2004, p.17). Comments such as these acknowledge that an ultimate understanding of the aetiological pathways which link socioeconomic position and health may demand a synthesis of the various hypotheses which have been developed. Under this view, the proposed explanations are arguably best regarded not so much as mutual competitors, but rather as discrete elements contributing to some single unified framework whose structure remains as yet unclear. In fact, the likelihood that an effective explanation of health inequality would be multi-faceted, drawing together elements from multiple conceptual models, was recognised in the Black Report itself:-

> "...there can be little doubt that amongst all the evidence there is much that is convincingly explained in alternative terms: cultural, social selection and so on. Moreover, it may well be that different kinds of factors, or forms of explanation, apply more strongly, or more appropriately, to different stages of the life-cycle." (Townsend *et al*. 1992, p. 115)

***

The discussion in the present chapter has sought to present, in brief outline, the main theories which have been proposed to explain the near-ubiquitous phenomenon of social differences in health (which was itself considered in Chapter 1). In preparation for introducing the objectives of the present study, attention now turns to consider a specific factor - that of housing conditions and the residential environment - which has been widely recognised as exhibiting associations with both sides of the health inequality 'equation' (that is, with both SEP and health).

CHAPTER 3 INTRODUCTION (III) - HOUSING CONDITIONS AND HEALTH

3.1 Socioeconomic position, housing and health

As was indicated in Section 1.2, a range of factors which represent aspects of housing (including tenure, physical conditions and amenities) has been used to measure the material aspects of an individual's socioeconomic circumstances. Of particular interest in the present study are the physical attributes of the residential environment: "household conditions such as the presence of damp and condensation... ...are housing related indicators of material resources." (Galobardes *et al*. 2006a, p. 9). Such a statement expresses the broad (and intuitively plausible) expectation that poorer people will, in the main, live in homes which are inferior in one or more respects (e.g. by being damper, colder or more crowded) to the dwellings occupied by those who are better off. An explicit assertion of this expectation is given by Hopton and colleagues: "Poor quality housing is more likely to be inhabited by those living on lower income, rather than higher income earners." (Hopton *et al*. 2003, p. 20). A similar view is evident in a report by Shaw and colleagues of a study which compared two 'extreme' areas of the UK (defined as groups of Parliamentary constituencies which exhibited, respectively, high mortality coexisting with low SEP, and low mortality accompanied by high SEP): "It is likely that in the 'best health' areas [i.e. those characterised by high SEP] the rooms were larger, lighter and less damp." (Shaw *et al*. 1999, p. 58).

While housing conditions may reflect material status, they are also widely acknowledged as an important predictor or determinant of health. There is thus an implied network of relationships among these three factors (SEP, housing conditions and health) which has been recognised by commentators. For example, Fuller-Thomson and colleagues have asserted: "Socio-economic status (SES) plays a central role in our efforts to understand the connection between housing and health." (Fuller-Thomson *et al*. 2000, p. 126). A similar sentiment has been expressed by Shaw: "housing remains a key social determinant of health and a central component of the relationship between poverty and health." (Shaw 2004, p.413). Clearly, relationships among these factors could take a number of forms. For example, both SEP and housing may exert independent effects on health. Alternatively, SEP might influence health through a mediating effect of housing conditions. The latter is, broadly speaking, the hypothesis suggested in a previously-cited study by Ellaway & Macintyre: "In this model income or wealth enables people to buy homes, and homes which are owned tend to have less noise, damp and other hazards." (Ellaway & Macintyre 1998, p. 149). While this

hypothesis is plausible (and is indeed closely related to that which underlies the present study), uncertainty over how SEP, housing and health may be mutually inter-related remains: "A large gap still exists in our knowledge about the links and pathways between housing, socio-economic status and health status." (Fuller-Thomson *et al.* 2000, p. 109). The difficulty of establishing the *direction* of the relationships involved in this three-way nexus (that is, distinguishing between cause and effect) is highlighted by Shaw:-

> "Housing, health and poverty are still empirically related and conceptually interconnected. We often become tangled in circular explanations as we attribute the effects of poverty on health to poor housing, and the effects of poor housing on health to general poverty." (Shaw 2004, p.414)

In preparation for declaring (in Chapter 4) a hypothesis which posits a specific pattern of causal association among these factors, the present chapter complements the earlier consideration of relationships between SEP and health (Chapter 1) by discussing associations between housing conditions and health. The discussion of these latter associations is more detailed than that of the SEP / health relationship, for the following reason. While the near-ubiquity of social inequality in health is almost universally accepted, there is less certainty about the extent to which housing conditions exert a causal influence on health. The present chapter will argue that despite the size of the research effort which has been devoted to the topic, a surprising degree of uncertainty remains over the nature of links between housing conditions and health. The development of this argument requires that the evidence base be examined in greater detail than was the case for the largely uncontested premise that health is socially patterned.

3.2 Housing conditions and health: general comments on the evidence base

An extensive body of research, assembled over many years, has sought to clarify the associations between domestic residential conditions and human health. The literature generated by this research has been the subject of a number of reviews (e.g. Fuller-Thomson *et al.*, 2000; Krieger & Higgins, 2002; Hopton *et al.*, 2003; Shaw, 2004). Further reviews have concentrated on specific aspects of the home environment e.g. dampness and mould (Peat *et al.*, 1998) and indoor air quality (Institute for Environment and Health [IEH], 1996; IEH, 2001), and on the relationship of housing to particular aspects of health e.g. housing and asthma (Richardson *et al.*, 2005).

The evidence base devoted to housing-health relationships is sizeable; however, the quality of much of that evidence is such that, despite the volume of research undertaken, the links between housing conditions and health remain imperfectly understood. Limitations of the evidence base have been acknowledged by commentators. The author of one fairly recent review stated:-

> "...the evidence base concerning the direct effect of housing on health is not as substantial as we might expect… …Studies that relate housing to specific health outcomes tend to be small-scale, often reporting small effect sizes, and sometimes with conflicting results; in some ways the evidence base can be characterized as piecemeal." (Shaw 2004, pp. 402-403).

Similar views are expressed by the authors of another review: "Research into the relationship between housing and health has frequently been narrowly focused, fragmented, and of marginal practical relevance to either housing or health policy." (Fuller-Thomson *et al.* 2000, p. 109).

One notable limitation of the evidence base is that little of it has been generated by intervention or experimental studies. Such studies provide what is potentially the most robust class of evidence, a fact recognised in the health domain by the long-acknowledged supremacy of the clinical trial in medical research. For obvious practical and ethical reasons, it would be expected that intervention studies in housing and health (that is, investigations in which some aspect of the residential environment is materially altered, and the effects on health observed) are uncommon. Yet even when such expectations are entertained, the scarcity (and indifferent quality) of such studies is genuinely surprising. A review by Thomson *et al.* investigated studies dating from 1887 onwards (in any language and format) and identified only 18 completed primary intervention studies, concluding "We found few studies examining the effects of housing improvements on health, and the quality of the studies identified was generally poor." (Thomson *et al.* 2001, p. 189). Consequently, much of the evidence relating to the health effects of housing is drawn from observational studies, which offer a less robust inferential base. Indeed, much of the evidence has been provided by cross-sectional investigations (which largely preclude any form of reliable causal inference). However, although small, the number of intervention studies in the field of housing and health continues to grow. Recent examples include studies reported by Howden-Chapman and colleagues (2007) and by Walker *et al*. (2009). An updated version of the previously-cited review by Thomson and colleagues is currently awaiting publication, and while this may not be cited in detail at present, these authors report that the number of

intervention studies has increased since their earlier report, and that the quality of recent studies has improved (Thomson *et al*., 2009 [personal communication of 'in press' publication]).

It thus appears that while much research has investigated relationships between the domestic environment and health, the value of the evidence gained from that research (in terms of being able to account convincingly for how housing and health are related) is lower than might be expected. Consequently, it is generally not the case that specific housing conditions are linked unequivocably, via fully understood causal mechanisms, with health. Less formally, it might be said that the commonly-held (and intuitively attractive) view that 'bad housing is bad for you' is not in fact supported by fully convincing research evidence. This view, while completely understandable, is held not only by laymen. For example, a recent past Chairman of Council of the Royal College of General Practitioners, writing in 2002, made reference to "the very considerable evidence base for the very clear link between housing and health". (Haslam 2003 [Foreword to Gill & de Wildt, 2003], p. v). While this statement correctly describes the evidence base (it is indeed considerable), its characterisation of the housing / health relationship as 'very clear' is more contentious (as demonstrated by the quotations from Fuller-Thomson and Shaw which were cited earlier). However, while recognising the limitations of the evidence (and the paucity of robustly demonstrated conclusions), some elucidation of the housing / health relationship has undeniably been achieved. Commentators recognise this progress, asserting that "the mechanisms through which specific aspects of housing affect health are extremely complicated, but they do exist. Researchers have made a great deal of progress in clarifying some of these mechanisms." (Fuller-Thomson *et al.* 2000, p. 109). In a similar vein, Shaw (having previously highlighted the piecemeal nature of the evidence base) states: "when amalgamated, the sum of the extensive range of ways in which housing is related to health is quite considerable." (Shaw 2004, p. 403). Overall, the current state of knowledge in the field of housing and health may be characterised as being neither in its infancy, nor developed to such a state of advancement that only minor details remain to be clarified, and small gaps in understanding filled in. Progress in elucidating the housing – health relationship continues to be made, including the identification of how housing may influence health not only directly, but via more complex routes which involve interactions with other factors. A good example of such an interaction mechanism is provided by the 'inverse housing law' identified by Blane and Mitchell (Blane *et al.*, 2000; Mitchell *et al.*, 2002). These authors demonstrated a mismatch between housing quality and local climate in the UK, such that some of the poorest housing is found in areas which experience the harshest climatic

conditions. This interaction between housing and environment was found to exhibit associations with both lung function and the risk of hypertension.

***

Of the literature reporting studies into housing and health, a considerable proportion relates to the associations between residential conditions and respiratory disorders. Unlike much research in the area, this work has indicated convincingly that certain associations do exist between specific aspects of the home environment and respiratory disease: "The main health outcome shown to be related to housing is that of respiratory health, measured by the presence of respiratory disease or by lung function." (Shaw 2004, p. 403). As will be indicated in Chapter 4, respiratory health is one of the two main outcome areas featured in this study. Reflecting this, the research evidence applicable to relationships between residential conditions and respiratory disease is now discussed.

3.3 Housing conditions and respiratory disorders

3.3.1 Representation of housing conditions in the research context

Before proceeding to discuss the literature devoted to housing conditions and respiratory health, some observations relating to how the concept of 'housing conditions' is realised in the research context will be helpful. Most research in the domain of housing conditions and health has focussed on specific aspects of the domestic environment e.g. the presence of dampness (Hopton *et al*., 2003). Few studies, if any, have attempted to work with a fully comprehensive definition of housing conditions; that is, a definition which includes concurrently all individual dimensions of housing conditions (e.g. the simultaneous presence of damp / mould, noise, crowding, indoor air pollution etc). Using such a comprehensive definition in a research context would be highly challenging, due to the number of individual dimensions of housing conditions which would have to be measured concurrently. To illustrate, a recent legislative definition of housing quality, currently in force in the UK, identifies no fewer than 29 types of housing 'hazard', each associated with a specific aspect of the home environment (Office of the Deputy Prime Minister, 2006). Faced with such complexity, it is not surprising that "Studies providing original data on the relationship [between housing and health], which is the vast majority of the literature, focus on very

specific physical, chemical, and biological exposures with a known or suspected effect on health within the house." (Fuller-Thomson *et al.* 2000, p. 109).

3.3.2 Housing conditions and asthma – the house dust mite Dermatophagoides pteronyssinus

The specificity of research effort identified by Hopton *et al*. and by Fuller-Thomson *et al.* is clearly evident in studies of how housing relates to respiratory health. Specific aspects of the home environment investigated for evidence of association with the respiratory health of occupants include hygrothermal conditions (cold, humidity and dampness, whether considered jointly or treated as separate phenomena); mould and fungi; and house dust mites. Discussion of relationships between these specific housing-related factors and respiratory health begins with the last-named.

The existence of a causal association between the presence of dust mites in the home and respiratory disease (specifically, the clinical entity termed 'asthma') has been hypothesised for many years[7]. A causal relationship between dust mites (especially the species *Dermatophagoides pteronyssinus*) and the allergen causing atopic asthma and rhinitis was proposed in the 1960s (Voorhorst *et al.*, 1967), since when many studies have investigated the postulated link between mites and allergic diseases, notably asthma. Evidence which may be considered supportive of a causal role for mites in the aetiology of asthma has been discussed by Platts-Mills *et al.* (1989), who cite studies conducted in many parts of the world. Fuller-Thomson *et al*., in a review of the housing / health relationship, argue that the evidence for a causal effect of mites on asthma should be classified as either 'Definitive' (defined as "numerous well-designed studies showing the effect, most or all causal criteria met, essentially complete agreement among experts that a health effect exists") or 'Strong' ("some well-designed studies showing the effect, most causal criteria met, a preponderance of opinion among experts that a health effect exists"). (Fuller-Thomson *et al*. 2000, pp. 123-124). More recent reviews by Richardson *et al*. (2005) and by Platts-Mills and colleagues (2009) also indicate broad acceptance of a causal relationship between mites and asthma. The former conclude "There is currently only reasonable evidence for one causative factor

[7] No attempt is made here to provide a definition of the term 'asthma', since no single universally agreed definition exists. To illustrate, Samet (1987) cites four differing definitions created by, respectively, the Ciba Guest Symposium; the American Thoracic Society; the American College of Chest Physicians; and the World Health Organisation. The absence of a single accepted definition presents challenges in the conduct of research into the aetiology of asthma: "The absence of a universally accepted definition of asthma makes it especially difficult to arrive at a consistent operational definition for epidemiologic studies." (National Academy of Sciences Institute of Medicine [NAS] 2000, p. 23)

for asthma in the indoor environment and that is house dust mite allergen." (Richardson *et al*. 2005, p. 328). In a similar vein, Platts-Mills *et al*. assert that

> "...taken together, the evidence provides a compelling case that perennial mite exposure is *(A)* the single most important cause of sensitization in patients with asthma worldwide, and *(B)* the cause of asthma in subjects with mite allergy." (Platts-Mills *et al*. 2009, p. 112)

The mechanisms via which dust mites are hypothesised to exert influence as an asthma-inducing agent, and the associated respiratory response in humans, have been described (Jones, 1998; NAS, 2000). The former paper forms the basis of statements presented in the remainder of this paragraph, no further specific citation of this source being made in the paragraph. Citing a range of authorities, Jones reports that faecal particles of the mite *Dermatophagoides pteronyssinus* are the principal source of antigens in house dust. The species thrives in moist, warm rooms: the optimal temperature for the organism is in the range 20-30^{o}C, the optimal relative humidity falling between 70% and 80%. Faecal particles produced by this mite are encased in a coating of intestinal enzymes, and it is a protein within these enzymes which is the primary allergen (known as *Der p* I). It is estimated that exposure to mite allergen may trigger attacks in up to 85% of asthmatics. The response to exposure is generally in two stages, the first involving a short episode of airway obstruction which normally resolves within one to two hours, while the second stage (which may develop some hours after exposure) usually involves inflammation of the airways. There is evidence that as well as being a trigger of attacks in asthmatics, exposure to mite allergen (especially in infants) may contribute to the onset of asthma itself.

The hypothesis that dust mites play a causal role in the aetiology of asthma (whether in initiating the condition in previously disease-free individuals, or by exacerbating existing airway disease) has enjoyed wide acceptance. This may be illustrated by quoting from a further selection of studies:-

> "An association between house mites and bronchial asthma has been recognised epidemiologically for many years." (Whyte & Flenley 1986, p. 89)

> "The role of house dust mite allergens in the provocation of atopic asthma has become established since 1967." (Dorward *et al*. 1988, p. 98)

> "As well as being a major trigger of attacks in asthmatics, there is also evidence that exposure to mite allergen, particularly amongst infants, may be an important factor in inducing the onset of asthma itself." (Jones 1998, p. 756)

> "House-dust mite exposure is one of the best documented environmental causes of asthma". (Peat *et al.* 1998, p. 121).

Such statements demonstrate that the postulated link between mites and asthma is regarded by many commentators as highly plausible. Indeed, the initial declaration by Voorhorst and colleagues that dust mites are the major source of allergen in house dust has been described as "one of the most important events in the history of allergic disease." (Arlian & Platts-Mills 2001, p. S406). However, a second body of evidence exists which is consistent with an alternative view: that mites are 'innocent bystanders'[8] in the processes involved in the development and / or exacerbation of asthma. Expressions of this alternative opinion include a review of environmental influences on asthma by Strachan, which concludes "the importance of mite allergen exposure in determining the prevalence and severity of asthma remains controversial." (Strachan 2000, p. 874). In the same vein, a review by Pearce *et al*. argues that "the evidence linking allergen exposure to asthma is weak". (Pearce *et al*. 2000, p. 430). In the latter case, the authors acknowledge the role of mite allergen exposure in secondary causation (i.e. triggering asthma attacks in already-sensitised subjects), but query the influence of allergens (including *Der p* I) on the risk of developing asthma. A detailed defence of the 'innocent bystander' school of thought is presented in a very recent review by von Hertzen and Haahtela (2009a). These investigators argue that "Novel prospective data, data from avoidance studies, as well as studies from farming and other microbe-rich environments, do not lend support to the dogma that dust mites are causally associated with asthma and related symptoms."

The examination by von Hertzen and Haahtela of evidence from avoidance studies (that is, studies in which subjects' level of exposure to mites is experimentally reduced, and the effects of the intervention on health assessed) is of particular interest, because such studies arguably provide an intuitively appealing method for testing the hypothesis that mite exposure influences asthma. Simply stated, "If dust mites are causally linked to asthma, it is logical to assume that a reduction in mite allergen level would improve the condition of a patient with asthma." (von Hertzen and Haahtela 2009b, p. 121). While some studies have suggested beneficial effects attributable to a deliberately-induced reduction in exposure to mites (e.g. Walshaw & Evans, 1986; Dorward *et al.*, 1988; Korsgaard & Dahl, 1994), a systematic review of randomised trials which assessed the effect of reducing exposure to mite antigens in the homes of individuals with mite-sensitive asthma did not find convincing

[8] The phrase is used in a very recent review paper by von Hertzen and Haahtela (2009a), which is considered shortly.

evidence of clinical benefit. Gotzsche & Johansen (2008) examined 54 trials, and conclude "We were unable to demonstrate any clinical benefit to mite-sensitive patients with asthma of measures designed to reduce mite exposure." It may be argued that the failure to detect a beneficial effect does not directly support the 'innocent bystander' view of dust mites, and that the negative findings of Gotzsche & Johansen may reflect, *inter alia*, inadequacies in the trials which were reviewed. However, these authors explicitly state that no effect was found "Despite the fact that many trials were of poor quality *and would be expected to exaggerate the reported effect*." [emphasis added] (Gotzsche & Johansen 2008, p. 646).

In addition to the generally negative results from experimental avoidance studies, von Hertzen and Haahtela argue that the 'innocent bystander' theory is supported by two other major strands of evidence. First, it has been demonstrated that certain living environments (specifically, "farming and simple living conditions") are associated with elevated exposure to mites, but with reduced risk of asthma. Second, a group of recent prospective studies (reported from the year 2000 onwards) of mite exposure and asthma among children do not generally support the hypothesis that increased exposure to dust mites in early life is associated with a higher risk of asthma.

It thus appears that a causal role for house-dust mites in the development and / or exacerbation of asthma, while widely viewed as plausible, is not unequivocably supported by research evidence. The previously-cited reviews by Platts-Mills and colleagues (2009) and by von Hertzen and Haahtela (2009a) may, in view of their recent provenance, be regarded as contemporary statements of (respectively) the cases for and against the existence of such a causal link[9]. On this basis, it may reasonably be concluded that a robust causal relationship between this specific feature of the residential environment (i.e. the presence of mites) and asthma remains at present unproven, and the subject of live controversy.

***

In addition to the postulated link between dust mites and asthma, other aspects of the domestic environment have been implicated in respiratory disorders. The evidence for some of these other putative effects of housing on respiratory health is now considered.

[9] Indeed, they were explicitly presented as 'pro' and 'con' views of the mite / asthma controversy in the same issue of the *American Journal of Respiratory and Critical Care Medicine*.

3.3.3 Housing conditions and respiratory disorders - mould

Another biological housing factor (as distinct from inanimate aspects of the home environment such as temperature and humidity) hypothesised to influence health is the presence of moulds and / or fungi[10]. The influence of these organisms on respiratory health or symptoms has been extensively researched, and evidence of associations obtained. For example, mould was found to be associated with wheezing and nocturnal cough in children (Strachan and Elton, 1986), and with sore throat in children (Martin *et al*., 1987). The literature relating mould to respiratory health has been the subject of a number of reviews, some of which are now considered.

Verhoeff & Burge (1997) identified nine population-based studies which examined associations between exposure to fungi in the home environment and the health of occupants; the outcomes examined related almost entirely to respiratory disease. From their review of the evidence, Verhoeff and Burge concluded "Fungi do contribute to allergic disease, and the extent of their involvement is probably greater than is indicated by the available clinical and epidemiological studies." (Verhoeff & Burge 1997, p. 552). However, the authors qualified their conclusions by acknowledging limitations of the evidence base, specifically the cross-sectional designs used in the studies examined (which limits the robustness of causal inferences), and "inconsistency and inadequate validation of the measures used to evaluate exposure and health effects." (Verhoeff & Burge 1997, p. 544). A review by Peat and colleagues highlighted a number of reported associations between dampness and mould[11] and wheeze and / or cough, but acknowledged that the estimated increased risk of these symptoms was fairly small. The conclusions of their review are summarised thus: "Clearly, most people are exposed to… …a wide range of mould spores in their home, but whether certain home characteristics lead to increased allergen exposure and, as a consequence, to respiratory health effects remain unclear." (Peat *et al*. 1998, p. 125). These authors also recognised methodological limitations to the studies examined, again identifying the inappropriateness of drawing causal inferences from cross-sectional designs, and the difficulty of measuring exposure to the hazard of interest. In connection with the latter, Peat *et al*. made the uncompromising assertion that "There are no practical and

[10] The terms 'mould(s)' and 'fungi' are frequently used interchangeably in the literature on housing and health e.g. "dampness and fungal (mold) problems are present in 20% to 50% of modern homes." (Verhoeff & Burge 1997, p. 544). This practice is generally adhered to in this thesis, references to 'mould(s)' or 'fungi' implying 'mould(s) and / or fungi'.

[11] To some extent this review treated damp and mould as a single factor, and this practice is in fact fairly common in the literature (see Section 3.3.4).

accurate methods with which to monitor mould exposure." (Peat *et al*. 1998, p. 123). Fuller-Thomson *et al*. assessed the strength of evidence relating dampness and mould to three specific areas of respiratory health: asthma, respiratory symptoms, and respiratory tract infections. For each of these three associations, the strength of evidence was classified as 'Possible', defined thus: "small number of studies showing the effect, some or few causal criteria met, no consensus among experts that a health effect exists." (Fuller-Thomson *et al*. 2000, p. 123). It is noteworthy that these authors were more cautious in their evalution of these associations than in their assessment (cited earlier) of the evidence relating to links between dust mites and asthma, which were classified as either 'Definitive' or 'Strong'. A review by Strachan concluded that studies investigating the relationship between home dampness or domestic mould growth and asthma had yielded fairly consistent results, indicating that wheeze was around twice as likely to occur in homes reported to be mouldy (Strachan, 2000). However, methodological limitations were again identified, the author commenting that assessment of both exposure and disease in such studies was generally self-reported (via questionnaires). Studies which featured objective assessment of home conditions and of symptomatology were fewer in number, and in the main demonstrated little relationship between mould and the presence of respiratory disease. This raised the possibility that some of the associations observed in questionnaire-based studies might be due to reports of symptoms being artificially increased by respondents' awareness of mould in the home, or *vice versa*.

The possible influence of reporting bias had been identified earlier by Strachan in connection with his own studies, defining the potential problem thus: "Parents of symptomatic children may be more aware of potentially adverse environmental circumstances or parents who perceive their housing to be unsatisfactory may report symptoms of a different degree of severity than others." (Strachan & Elton 1986, p. 141). A subsequent study by the same author (Strachan, 1988) concluded that reporting bias explained a substantial part of the association observed in the study between damp or mouldy housing and wheeze. From this was drawn the more general (and potentially highly important) conclusion that "further studies of this relation [i.e. that between mould and respiratory symptoms] are unlikely to be valid if they rely solely on information from questionnaires." (Strachan 1988, p. 1226). Taken in conjunction with the conclusion of Peat *et al*. (quoted earlier) regarding the difficulty (indeed, the implied impossibility) of obtaining accurate objective measures of mould exposure, Strachan's assertion assumes a disquieting significance. If the hazard cannot be accurately characterised (Peat *et al*.), and the assessment of symptoms (when

based, as was frequently the case in studies in this area, on self-report) is unreliable (Strachan), conclusions cannot be considered robust and the reliability of a large portion of the evidence base relating to mould and respiratory health is potentially undermined. In an extension of the study just cited, Strachan and colleagues investigated the issue of differential reporting further via the approach of obtaining both subjective assessments and objective measurements of mould in the homes of subjects (Strachan *et al.*, 1990). The authors concluded that questionnaire reports of mould in the home may be a poor indicator of actual exposure to airborne spores, presenting this study as "confirmation of differential reporting of both symptoms and housing conditions in questionnaire data." (Strachan *et al*. 1990, p. 386).

A more recent review by Richardson *et al*. (2005) identified conflicting conclusions in the literature in respect of associations between the presence of residential mould and respiratory disease. For example, a review by the Medical Research Council Institute for Environment and Health is cited by Richardson *et al*. as identifying "insufficient evidence" for a relationship between mould and respiratory health (Richardson *et al*. 2005, p. 330). Reference to the original source confirms this interpretation: "Although there is consistent evidence of a link between damp and mouldy housing and reports of respiratory symptoms in children, the few epidemiological studies conducted to date show no convincing specific association between exposure to indoor airborne fungi and respiratory disease." (IEH 2001, p. 17). Conversely, Richardson *et al*. cite another review by the Institute of Medicine of the US National Academy of Sciences as reporting evidence of exacerbation of respiratory disease (specifically, of asthma) by mould. However, reference to the original source (NAS, 2000) actually reveals a slightly more nuanced expression of conclusions. The effects of fungi in this review are considered in a chapter entitled 'Indoor Dampness and Asthma', reflecting a previously-highlighted tendency in the literature to regard dampness and fungi / mould as manifestations of a single underlying phenomenon (see Footnote 11 above). The NAS review concludes:-

> "1. Damp conditions are associated with the existence of doctor-diagnosed asthma and with the presence of symptoms considered to reflect asthma… …2. Symptom prevalence among asthmatics is also related to home dampness indicators… …3. The factors related to dampness that actually lead to the development of disease and to disease exacerbation are not yet confirmed, but probably relate to dust mite and fungal allergens." (NAS 2000, p. 310).

The material discussed above suggests that, in contrast to the well-defined, plausible and widely accepted link between house dust mites and asthma, there is a lack of consistency in

the evidence relating residential mould to respiratory disease (including asthma). As a result, there is a corresponding absence of consensus on the nature of causal links between mould and respiratory ill-health, despite the scale of the related evidence base. The key reasons for this lack of consensus are (a) the dominance of cross-sectional studies (from which causal pathways cannot be confirmed) in the field, and (b) the difficulty of accurately quantifying both exposure to mould in the home and the health status of study participants. Three important points which potentially complicate the interpretation of study findings, and so contribute to the lack of definitive conclusions, were made by Strachan *et al.* (1990). First, although mould growth on walls has been found to increase the average indoor spore burden, even these elevated concentrations may be modest compared to outdoor spore levels: "...only three of the [indoor] samples in our study approached the levels of 10 000 – 50 000 CFU/m^3 that are typical of a summer garden." (Strachan *et al.* 1990, p. 385). From this, it may reasonably be inferred that exposure to spores indoors represents only a proportion (possibly small) of the total spore burden to which an individual is exposed. Second, many of the species of indoor mould identified in this study were similar to those normally found out of doors. From this, the authors argued that explicit mould growth in homes is unlikely to influence the quality of indoor fungal flora to an important extent. Finally, most of the species identified in homes in this study of Strachan and colleagues (including those species known to be highly allergenic) were isolated on at least one occasion in most homes. This suggests that most people are exposed to some extent to a wide range of fungi in the indoor air, posing challenging questions about possible causal links between spore exposure and respiratory disease. If most individuals are exposed to a common range of spores in the home (albeit varying in the level of exposure and in species mix), what factors lead to some individuals developing respiratory symptoms, while others do not? One answer to this question, certainly in the case of the clinical entity termed 'asthma', is that genetic (inherited) characteristics are also involved in the development of disease. That predisposition to asthma is an inheritable trait is widely accepted; however, the manifestation of disease is influenced by both inherited and environmental factors (NAS, 2000). This adds a further layer of difficulty in investigating relationships between housing conditions and asthma.

One further consideration which has hindered efforts to identify the influence of mould on respiratory health is the difficulty of isolating any specific effect attributable to mould from the effects of other aspects of the domestic environment. The challenge is delineated succinctly by Boardman: "For instance, with asthma what is the relative effect of mould, dust mites and passive smoking?" (Boardman 2000, p.5). This difficulty in separating out

the effects of individual co-existing features of housing conditions is a major challenge to research in the housing / health domain.

Overall, the evidence relating to links between residential mould and respiratory disease may be summarised as follows. Although the presence of mould in the home has been repeatedly demonstrated to exhibit associations with respiratory symptoms (especially among children[12]), methodological limitations attached to many studies in the field preclude reliable inferences which postulate a causal relationship. Because of this, it is argued that the case for the existence of a direct causal link between mould and respiratory illness cannot be regarded as proven at the present time.

***

Having considered links between house dust mites and asthma, and associations between residential mould and respiratory disease, the relationship of hygrothermal conditions in the home to respiratory health is now discussed.

3.3.4 Housing conditions and respiratory disorders – the hygrothermal environment

As previously indicated, there is a tendency in the literature on housing and health to treat 'damp and mould' in the home as a single entity. This approach is exemplified in a review by Peat *et al*. which "examines whether there is a direct or indirect relation between damp or mould in the home and respiratory health." (Peat *et al*. 1998, p. 120). Here, a biological agent (mould) is conflated with a non-organic, inanimate attribute of the physical environment. This practice is in fact fairly common in the literature, and reflects (perhaps unconsciously) established knowledge that the two factors of damp (or, more accurately, moisture levels) and mould are related: the availability of sufficient moisture is normally the critical factor determining whether mould will grow in a dwelling (Oreszczyn & Pretlove, 2000). A similar process of conflation may be observed in connection with an occasional tendency in the literature to treat two different components of the hygrothermal environment

[12] The focus of many studies on the respiratory health of children is itself arguably a limitation of the evidence base. Although the study of child cases avoids possible confounding influences (such as exposure to occupational respiratory hazards, and the respiratory consequences of active [though not passive] smoking), children are particularly vulnerable to housing-related respiratory hazards because of the amount of time they spend in the home. In consequence, even where associations between mould and respiratory symptoms have been identified in children, such findings may not be reliably generalisable to adult populations.

(moisture levels and temperature) as a single construct. For example, one paper contains the assertion that "many symptoms might be related to cold, damp conditions" (Strachan and Sanders 1989, p. 7), while another refers to "relationships between cold, damp housing and poor health." (Collins 2000, p. 39). As with the conceptual blurring of mould and dampness, the reasons for this combination of two different dimensions of the indoor environment into a single factor are understandable, in that damp houses tend to be colder than dry dwellings (Hopton, 2003).

This elasticity of conceptual boundaries illustrates a major challenge facing much research into the relationships between housing and health. As Boardman states "the relative effects of cold, damp and mouldy living conditions are difficult to disentangle as they are co-related." (Boardman 2000, p.6). In fact, such co-relation is not restricted to the three-way linkage given by Boardman. Hygrothermal conditions are also related to the proliferation of house dust mites, research having established that the main dust mite species implicated in asthma (*Dermatophagoides pteronyssinus*) proliferates at relative humidity levels above 73% and temperatures above 25^{o}C (Oreszczyn & Pretlove, 2000). Furthermore, dust mites and mould are themselves subject to an additional important ecological relationship in that the main food of mites is human skin scales (Howieson & Lawson, 2000), and mites rely on mould to break down the skin particles and make them digestible (Oreszczyn & Pretlove, 2000). Yet another relationship to be considered is that linking temperatures and moisture levels in the home to ventilation. Ventilation rates have an important impact on the dispersion and diffusion of water vapour (Howieson *et al.*, 2003), and are thus a determinant of moisture levels in the home. Clearly, inter-relations between the various potential determinants of respiratory health (including hygrothermal conditions, which are the topic of the present section) are complex.

The above considerations suggest that temperature and / or moisture levels in the home may potentially impact the respiratory health of occupants by acting to promote (or hinder) the growth of possibly harmful organisms. This may be conceptualised as an indirect effect, in that it operates via the organism(s) whose development is promoted (or hindered). However, the possibility of a direct effect must also be considered; that is, the scope for specific temperature and / or moisture level ranges to generate direct physiological effects independent of those arising from the mediating influence of organisms such as mites and mould. Consideration of the evidence base indicates that, for respiratory health, the two dimensions which comprise hygrothermal conditions - temperature (specifically low temperature) and moisture levels (particularly high levels) – map fairly closely onto the

direct / indirect distinction just discussed. Most research on the respiratory health impacts of low indoor temperature has focused on its possible direct physiological effects, while investigations into the influence of moisture levels have concentrated mainly on the role of moisture as an indirect determinant of respiratory health, operating by promoting the proliferation of mould and / or dust mites. Many researchers indicate acceptance of the hypothesis that indirect routes provide the most plausible explanations for observed associations between elevated moisture levels in the home and ill-health. For example, Strachan and Sanders state "On biological grounds, the most plausible links between dampness and respiratory disease implicate functional abnormalities of the airways, resulting from increased exposure to airborne allergens." (Strachan and Sanders 1989, p. 13). Similarly, Evans *et al*. assert that "the association between damp housing and some aspects of health is biologically plausible through the effects of house dust mites and moulds". (Evans *et al*. 2000, p. 677). This view is dominant in the literature, there being little suggestion that moist air in the home exerts a direct adverse effect (i.e. one not mediated via biological agents) on respiratory health.

Attention now turns to evidence of relationships between low indoor temperatures and respiratory health. In fact, the literature on this particular facet of the housing / health relationship is rather sparse: "Few studies have directly examined the relationship between cold housing and health." (Hopton *et al*. 2003, p. 33). In the same vein, Collins observes that "Most studies have focused on damp and mouldy living conditions rather than cold house temperatures." (Collins 2000, p.39). Low temperatures are known to produce measurable physiological changes in the human respiratory tract through cooling and drying of the mucosal surfaces. Exposure to cold is linked with impaired lung function (as measured by forced expiratory volume), and cold is a triggering factor for bronchoconstriction in individuals suffering from asthma or chronic obstructive pulmonary disease (COPD) (Collins, 2000). In addition, low temperature (acting via the suppression of immune responses by stress hormones during cold exposure) is considered likely to reduce resistance to respiratory infection (The Eurowinter Group, 1997). However, although such biologically plausible mechanisms exist to explain adverse direct health effects of low temperatures in the home, "it is methodologically very difficult to demonstrate a definite link between home temperatures and specific health outcomes." (Collins 2000, pp. 46-47).

One body of research which potentially provides insights into the relationship between cold housing and respiratory health relates to the phenomenon of excess winter mortality (that is,

higher death rates observed during the winter months relative to the remainder of the year)[13]. Elevated mortality during times of cold weather has been observed in locations as diverse as Israel, New Zealand, Bangladesh, Hawaii and Moscow (Gemmell *et al*., 2000). The phenomenon is also evident in many European countries, including (perhaps surprisingly) those located in Southern Europe such as Spain and Portugal (Healy, 2003). Evidence of excess winter mortality has been demonstrated for the UK, the number of excess winter deaths recorded in England and Wales during the period 1993 to 2000 ranging from 25,900 (in the winter of 1993/4) to 48,440 (1999/2000) (Johnson & Griffiths, 2003). Writing in 2000, Keatinge & Donaldson asserted that around 40,000 excess winter deaths were observed in Britain every year, the rate of excess mortality being the highest in Europe (Keatinge & Donaldson, 2000). Although deaths from a number of causes exhibit increases in winter, the main increase is in respiratory and cardiovascular mortality (Wilkinson *et. al*., 2000). Collins confirms that "Respiratory mortality contributes a significant proportion of excess winter deaths." (Collins 2000, p. 44). However, the contribution of the residential environment to total excess cold-related mortality continues to be debated. This is due in part to the difficulty of separating out the respective independent influences of outdoor and indoor conditions during periods of low temperatures. A number of studies have examined aspects of the associations between low indoor temperature, cold external conditions and mortality. While not providing definitive conclusions about these associations, some clarification has been achieved. A study by Keatinge demonstrated that substantial winter mortality was experienced even among individuals who enjoyed unrestricted home heating, but were exposed to cold conditions during outdoor excursions (Keatinge, 1986). This suggests that excess winter mortality is not purely a function or product of low temperatures in the home, but that the external environment is also implicated. Another study, which examined the relationship between cold exposure and winter mortality from a number of causes (including respiratory disease) in both warm and cold regions of Europe, found evidence which linked mortality with home heating independently of outdoor cold stress, and with outdoor cold stress independently of home heating (The Eurowinter Group, 1997). This study identified a number of specific associations, finding that high indices of cold-related mortality from respiratory disease were associated with low living-room temperatures, with limited bedroom heating, and with low proportions of people wearing thermally-protective clothing (such as hats and anoraks) outdoors in cold weather. Taken together, the studies by

[13] Much of the following material is also applicable to the later discussion of associations between housing conditions and cardiovascular disease (see Section 3.4).

Keatinge and The Eurowinter Group imply that both indoor and outdoor cold may exert adverse effects on mortality (including deaths attributable to respiratory disease).

One point of considerable interest highlighted by the Eurowinter Group study was that increases in all-cause mortality in cold weather were relatively greater in warmer regions than in colder regions. For example, the percentage increase in all-cause mortality per 1°C fall in temperature below 18°C was estimated to be 2.15% for Athens, as against 1.37% for London and 0.27% for South Finland. This finding suggests that absolute temperatures are not the sole determinant of levels of excess winter mortality. Support for this conclusion was given by a study conducted by Donaldson *et al*. in the Yekaterinburg region of Russia, where the mean winter temperature is -6.8°C, lower than in any part of Western Europe (Donaldson et al., 1998a). This study found that mortality from a range of causes (including respiratory disease) did not change until the mean daily temperature dropped below 0°C. The authors concluded that excess mortality was prevented by increasing the number of items of clothing worn, in combination with maintaining warmth in houses. Thus, a behavioural component of excess winter mortality was implied. A study by Clinch and Healy compared excess winter mortality in Ireland with that in Norway, positing a link between poorer housing standards (in terms of thermal efficiency and heating systems) and higher rates of excess winter mortality in the country with the milder climate (Ireland) (Clinch & Healy, 2000). This study found that relative excess winter mortality from respiratory disease in Ireland was 1.4 times the corresponding figure for Norway. The authors hypothesised that poorer housing standards in Ireland would allow falls in outdoor temperature to have a greater impact on indoor temperatures than would be the case in the superior housing conditions of Norway. A further study by Healy covering 14 European countries confirmed the 'paradox of excess winter mortality' i.e. the phenomenon that higher excess winter mortality rates "are generally found in less severe, milder winter climates where, all else equal, there should be less potential for cold strain and cold related mortality." (Healey 2003, p. 786). This study also considered domestic thermal efficiency standards, and concluded that those countries with the poorest housing (Portugal, Greece, Ireland, the UK) demonstrated the highest excess winter mortality.

On the basis of the evidence considered above, it may reasonably be concluded that there is some evidence of a link between cold conditions in the home and excess winter deaths from respiratory causes. However, evidence implicating the external environment indicates that low indoor temperatures are not the sole (or necessarily even the primary) factor responsible.

On this basis, it is difficult to assert that low indoor temperatures *in isolation* represent an influential risk factor for respiratory ill-health.

3.3.5 Housing conditions and respiratory disorders – vehicular emissions in the vicinity of the home

Although not strictly speaking an aspect of the residential environment, levels of motor vehicle emissions in the vicinity of the home have been linked with respiratory health impacts. This particular hazard is considered here because in the dataset used for the present study (see Section 5.3.2 in *Methods*), exposure to vehicle emissions (defined by the proximity of the subject's residence to major roads) is one of the factors used to assess exposure to air pollution in the home. Hundreds of different hydrocarbons have been identified in vehicle exhaust products, but the primary pollutants emitted by modern engines are carbon monoxide (CO), oxides of nitrogen (NO_x), and diesel particulates[14] (Utell *et al.*, 1994). Reviews by Dockery & Pope (1994) and Brunekreef *et al*. (1995) have concluded that particulates smaller than 10^{-6} metres (PM10) are associated with a range of adverse respiratory effects, including elevated mortality from respiratory causes, increased respiratory hospital admissions, greater reporting of lower respiratory symptoms and cough, and decreased peak flow. Another engine combustion product, nitrogen dioxide (NO_2), is linked with respiratory symptoms, lung function decrements and increased airway reactivity (Utell *et al.*, 1994). While outdoor and indoor concentrations of the primary vehicle-generated pollutants can vary greatly (Utell *et al.*, 1994), plausible evidence of adverse respiratory health effects on individuals living in areas characterised by high levels of vehicular traffic activity has been obtained. In one large multi-centre study (Ciccone *et al.*, 1998), high frequency of heavy goods vehicle traffic in the street of residence was associated with significantly increased risk of adverse respiratory outcomes (speech-limiting wheeze and persistent phlegm) in children experienced in the past 12 months. Significant associations were also found with a number of other respiratory symptoms, including nocturnal dry cough, morning chest tightness and persistent cough. Another study (Montnemery *et al*., 2003) found that living close to heavy traffic was significantly associated with recurrent or permanent nasal symptoms in adults, including nasal discharge, blocked nose and sneezing / itching. Vehicle emissions have also been associated with an increased risk of lung cancer. (Nyberg *et al*., 2000)

[14] Historically, the composition of vehicle emissions has altered over time, reflecting changes in legal emission control requirements. For example, prior to the elimination of lead compounds as anti-knocking additives, exhaust emissions included lead oxide particulates. (Utell et *al.*, 1994)

***

Having considered (in very brief outline) some of the ways in which the residential environment has been linked to respiratory ill-health, attention now turns to associations between housing conditions and the second main health outcome area considered in this study, namely disorders of the cardiovascular system.

3.4 Housing conditions and cardiovascular disorders

While the main health outcome shown to be related to housing is that of respiratory health (Shaw 2004), the domestic residential environment has also been implicated in cardiovascular ill-health. One housing-related factor which has been posited as exerting an influence on cardiovascular health is that of low indoor temperatures. While few studies have directly considered the effect of cold housing on health (Hopton *et al.*, 2003), research conducted in connection with the phenomenon of excess winter deaths (see Section 3.3.4) has demonstrated relationships between environmental cold and mortality from circulatory disease (e.g. Bull, 1973; Bainton *et al*., 1977; West, 1989; Khaw, 1995).

The proportion of excess deaths in cold weather attributable to cardiovascular disorders is substantial. Khaw, writing in 1995, cites a total of 40,000 excess winter deaths in the UK, asserting that "Most of these deaths are due to cardiovascular disease, predominantly heart attacks and strokes." (Khaw 1995, p. 337). Mercer (2003) states that cardiovascular disease accounts for the majority of excess winter deaths (up to 70% in some countries), while about half of the remaining excess mortality is due to respiratory disease. Recently published analysis by the UK Office for National Statistics indicates that circulatory disease was one of two main causes of excess winter deaths in England and Wales (the other being respiratory disease) during the winters of 2004/05, 2005/06 and 2006/07 (Brock, 2008). While evidence of associations between environmental cold and cardiovascular mortality is abundant, studies of seasonality-related circulatory disease morbidity are relatively rare (Mercer, 2003). Marchant *et al*. (1993), in a study of patients admitted to a coronary care unit with acute myocardial infarction (AMI), concluded that AMI is more common in winter and more common on colder days, independent of season. Spencer and co-workers (1998) observed seasonal variation in hospital admissions for AMI, the number of cases admitted in the winter months being 53% higher than in summer. A very large-scale 10-year longitudinal

study conducted by Danet and colleagues (1999) identified associations of atmospheric temperature with daily rates of both myocardial infarction and coronary death, rates of coronary events decreasing linearly with increasing atmospheric temperature.

Plausible physiological mechanisms for an adverse effect of cold on the cardiovascular system exist. Keatinge and colleagues (1984) identified a number of physiological responses to cold stress (increased platelet and red cell counts, blood viscosity, and arterial pressure) which offer a possible explanation for increases in both coronary and cerebral thrombosis in cold weather. Increases in arterial pressure in cold weather were also observed by Brennan and colleagues (1982), while Giaconi *et al*. (1988) found that ambulatory blood pressure measurements over a 24-hour period were higher in cold weather. Woodhouse *et al*. (1993) demonstrated that seasonal variation of blood pressure is heightened in older people, and argued that this finding may be partly responsible for increased cardiovascular disease mortality among the elderly in winter. These authors specifically posited a causal link between low indoor temperature and seasonal mortality, mediated by blood pressure:-

> "Although both indoor and outdoor temperatures were independently related to the seasonal variation of blood pressure, this association was quantitatively greatest for indoor temperature. The present data suggest that every 1°C decrease in living room temperature is associated with a 1.3-mmHg increase in SBP and a 0.6-mmHg increase in DBP. If blood pressure does mediate seasonal mortality from cardiovascular disease, might warmer living rooms decrease the number of winter deaths?" (Woodhouse *et al*. 1993, p. 1273)

Other risk factors for cardiovascular mortality have been found to exhibit seasonal variation. Stout and colleagues found that concentrations of the protein fibrinogen, which predicts the development of cardiovascular disease (Yarnell *et al.*, 1991) and is a risk factor for stroke (Wilhelmsen *et al.*, 1984), were higher in cold weather, and concluded that "The seasonal variation in plasma fibrinogen concentration is large enough to increase the risk of both myocardial infarction and stroke in winter." (Stout *et al*. 1991, p. 9). Cyclic seasonal variation in another risk factor for heart disease - circulating lipid levels - was identified by Gordon *et al*. (1988), the observed levels peaking in winter.

While much is known about the physiological pathways via which cold may affect cardiovascular health, the specific contribution of housing conditions to cold-related morbidity and mortality from CVD remains unclear. The problem of isolating the respective contributions of cold experienced in the outdoor environment and indoors is challenging, and has promoted continuing debate: "the debate about excess winter deaths is a debate about the

extent this is as a result of getting cold indoors and getting cold outside." (Boardman 2000, p. 5). In the specific context of cardiovascular disease, adherents of the latter view include Keatinge and colleagues, who have argued that "outdoor excursions rather than cold houses are the main cause of the arterial deaths that now cause most of the excess mortality in winter.[15]" (Keatinge *et al*. 1989, p. 76). Conversely, Wilkinson *et al*. assign greater importance to the role of the indoor thermal environment, while conceding that robust evidence for an effect of housing on excess winter deaths is limited:-

> "Though there is as yet little direct evidence that housing has an appreciable influence, extrapolation from what is known about patho-physiological mechanisms would suggest that cold homes are likely to have an appreciable impact on winter mortality". (Wilkinson *et al*. 2000, p. 26)

In summary, while clear evidence of elevated mortality from cardiovascular disease in cold conditions has been obtained, the contribution of the home environment to this phenomenon remains as yet imperfectly determined.

3.5 Housing conditions and health: summary

The present chapter has considered in brief outline some of the main relationships which are postulated to link the residential environment with two specific areas of health. Possibly contrary to lay belief, the evidence for a general adverse effect of 'poor housing' on health is, when considered *en bloc*, surprisingly weak. Some remarks on this topic were presented earlier (in Section 3.2), and attention is drawn to the previously-cited comments of Shaw and of Fuller-Thomson and colleagues which acknowledge the limitations of the evidence base (second paragraph of Section 3.2).

One widely (but not universally) accepted association is found in the area of respiratory disorders, where there is considerable support for the view that house dust mites are a causal factor for the onset and / or the exacerbation of asthma. However, as discussed in Section 3.3.2, evidence supportive of an alternative view exists; consequently, the mite / asthma link currently remains the subject of live controversy. Nonetheless, although considerable uncertainty exists over the extent to which *causal* mechanisms may be involved, the

[15] Proponents of the hypothesis that venturing outside in cold weather is the main risk factor for excess winter mortality have been described as the 'getting cold at bus stops' school. (Boardman, 2000)

evidence for a range of associations between aspects of the residential environment and health is substantial.

***

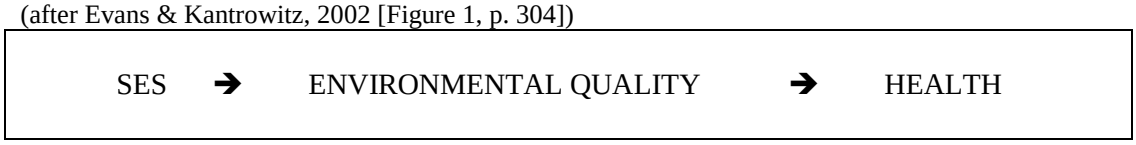
At this point, introduction of the themes which underlie the present study is complete. The next chapter draws on these themes to outline the motivation for the study, and describe the hypothesis which is tested.

CHAPTER 4 INTRODUCTION (IV) - A SPECIFIC HYPOTHESIS: HOUSING CONDITIONS AS A MEDIATOR OF SOCIAL INEQUALITIES IN CARDIO-RESPIRATORY HEALTH

4.1 Introduction: environmental mediation in health inequality

Thus far, three broad themes have been introduced: the near-ubiquity of social inequality in health (Chapter 1); proposed explanations for this phenomenon (Chapter 2); and associations between aspects of the physical residential environment and health (Chapter 3). A conceptual link connecting these three themes was considered in Section 3.2, concisely summarised in the observation that "housing remains a key social determinant of health and a central component of the relationship between poverty and health." (Shaw 2004, p.413). The concept of environmental mediation in the relationship between socioeconomic position and health is not restricted to housing conditions. The diverse range of such possible mediators was recognised in a review by Evans & Kantrowitz (2002), which was based on the premise that "Among several viable explanations for the ubiquitous SES-health gradient is differential exposure to environmental risk." (Evans & Kantrowitz 2002, p. 303). These authors worked with a very broad interpretation of environmental risk "…including hazardous wastes and other toxins, ambient and indoor air pollutants, water quality, ambient noise, residential crowding, housing quality, educational facilities, work environments, and neighborhood conditions." (p. 303). The conceptual model identified by Evans & Kantrowitz is illustrated in Figure 4.1.1.

FIGURE 4.1.1: Postulated relations linking socioeconomic status, environmental quality and health (after Evans & Kantrowitz, 2002 [Figure 1, p. 304])



In the representation of Figure 4.1.1, two features of are of interest. First, the arrows linking the three individual elements of the model explicitly indicate causation: socioeconomic status is posited as *influencing* environmental exposure, which in turn is assumed to exert an effect on health. Because it postulates a specific direction of effect, this model positively excludes 'reverse causation' explanations of health inequality such as the selection theory presented in the Black Report (see Section 2.1). A second, more subtle, aspect of the model shown in Figure 4.1.1 is that it incorporates a temporal dimension: both the hypothesised 'root' determinant of health (i.e. SES) and the postulated mediating factor (environmental risk) are potentially subject to variation over time. Thus, the individual's health status at a

specific point in life (say, *Y* years of age) may reflect varying patterns of exposure over time to both the determinant and the mediating influence. The principle that health is the product of varying experiences over time is the central premise of the lifecourse approach to health inequality (see Section 2.2); the model of Figure 4.1.1 therefore implies broad acceptance of the lifecourse paradigm.

Once the potentially dynamic (time-variant) nature of the exposures involved (that is, of SES and environmental risk) is recognised, the model of Figure 4.1.1 gives rise to three questions. First, in what ways may exposure to the social and environmental risk factors of interest vary over time? For example, do the majority of individuals experience relatively static patterns of exposure (characterised by few changes in socioeconomic position and / or environmental risk status), or are most people subject to frequent changes in these factors? Second, once time-related patterns of exposure have been identified, how do these patterns relate to health status at *Y* years of age? This is essentially the central question posed by the lifecourse approach to health inequality. In the context of the model under discussion here, specific questions include:-

- does an exposure (e.g. to social disadvantage) beginning at age *Y* years exert a different effect on health status from an exposure of corresponding duration which starts at (say) *Z* years $(Y \neq Z)$?

- do two exposures (of duration *a* and *b* years respectively; $a < b$), both beginning at age *Y* years, lead to differing health outcomes?

- is the effect of a single exposure of duration *a* years equal to that of multiple shorter exposures (of duration *b*, *c* and *d* years; $b + c + d = a$)?

Expressed more generally, the issue of interest is whether the timing, duration or degree of fragmentation of the exposure is associated with differences in health.

A final question arising from recognition that the exposures shown in Figure 4.1.1 are potentially time-variant quantities is essentially a methodological one. Consider a hypothetical pattern of exposure:-

AGE(years)	Y	Y+1	Y+2	Y+3	Y+4	Y+5... etc.
EXPOSED	yes	yes	no	yes	no	yes

Representing such patterns in a form suitable for use in statistical analysis and modelling (for example, to investigate the first two questions outlined above) presents certain challenges. Under a limited range of conditions, the task of representation is straightforward. If the exposure may be expressed as a simple exposed / not exposed contrast at each year, and if interest is restricted to the subject's total accumulated experience of the risk, all that is required is a simple summation of the number of years during which s/he was exposed. However, such an approach is obviously inadequate where either (a) the timing and / or degree of fragmentation of exposure (as distinct from a simple accumulated total) is of interest, or (b) exposure status cannot adequately be expressed as a simple binary contrast. An example of the latter arises when exposure to social disadvantage is expressed in terms of occupational class, which is a six-way ordinal scheme (see Section 3.1). Thus, investigating the first two questions involves addressing a third: how may possibly complex patterns of time-related experience be effectively represented in statistical analysis?

4.2 A specific instance of environmental mediation: SEP, housing conditions and cardio-respiratory health

The present study was designed to address all three of the questions outlined above within the context of a restricted version of the model proposed by Evans & Kantrowitz. For the study, the general model of Figure 4.1.1 was restricted in two respects. First, the generic concept of 'environmental quality' was reduced to a more specific set of exposures, namely those related to housing conditions and the indoor environment. This restriction largely reflected one of the student's previous research interests: relationships between housing conditions and health (Walker *et al.*, 2006; Walker *et al.*, 2009). A second restriction applied to the general model was that the health outcomes examined were limited to expressions of respiratory and cardiovascular health[16]. Although imposed partly by pragmatic considerations (specifically, availability of an appropriate data source), the selection of these health areas was shaped by the following considerations. As indicated earlier, respiratory health is the main health outcome shown to be related to housing (Shaw, 2004), and was therefore an obvious choice for an investigation of the Evans & Kantrowitz model in a housing context. Housing conditions have also been implicated in cardiovascular disorders (see Section 3.4), but specific motivation for considering circulatory ill-health in the present study was provided by the fact that cardiovascular disorders (particularly CVD

[16] In fact, a number of other health outcomes were examined by the study, but not in great depth – see Section 9.1 in *Methods.*

mortality) and their associated risk factors have frequently featured as outcomes in studies which hypothesise a lifecourse model of health inequality (e.g. Davey Smith *et al.*, 1997; Heslop *et al.*, 2001; Hallqvist *et al.*, 2004; Singh-Manoux *et al.*, 2004). Therefore, in investigating a specific path via which socioeconomic position over time might influence health, a concentration on cardiovascular health outcomes located the present study firmly within a well-established tradition.

***

After application of the restrictions described above, the specific form of the Evans and Kantrowitz model adopted for investigation in the study was as shown in Figure 4.2.1.

FIGURE 4.2.1: Hypothesised chain of association linking socioeconomic position with health via exposure to residential hazards.

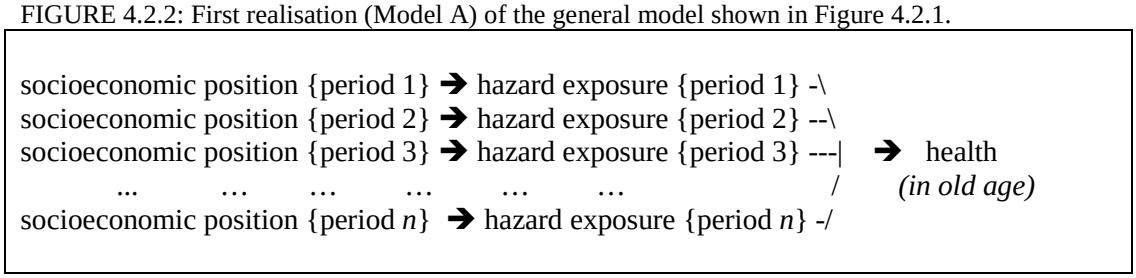
socioeconomic position ➔ exposure to residential hazards ➔ cardio-respiratory health *(assessed in old age)*

For the general model of Figure 4.2.1 to be valid, three sets of associations must be demonstrated. First, an overall 'end-to-end' relationship between socioeconomic position and health must be shown to exist. Second, evidence that SEP is linked to exposure to residential hazards (the leftmost association in Figure 4.2.1) must be identified. Finally, a relationship between exposure to these hazards and the health outcomes of interest (the rightmost association) needs to be demonstrated. The last two requirements were recognised by Evans and Kantrowitz in the review cited earlier, which identified "…two necessary prerequisites for this model to be valid - namely that socioeconomic status (SES) is associated with environmental quality and, in turn, that environmental quality affects health." (Evans & Kantrowitz 2002, p. 303)

The chain of associations shown in Figure 4.2.1 may be translated into a number of more specific forms, and three such variants were examined during the course of the study. The objective of assessing multiple variants of the general model was to allow both of the main 'strands' of the lifecourse approach to health inequality (i.e. critical period and accumulation of risk; see Section 2.2) to be tested. The first of three realisations of the general framework shown in Figure 4.2.1 was based on the premise that an individual's socioeconomic position at a specific stage in the lifecourse influences her / his exposure to housing hazards, either

contemporaneously or (if a 'lag' effect is involved) at some later point. This specific model (hereafter referred to as Model A) is illustrated in Figure 4.2.2.

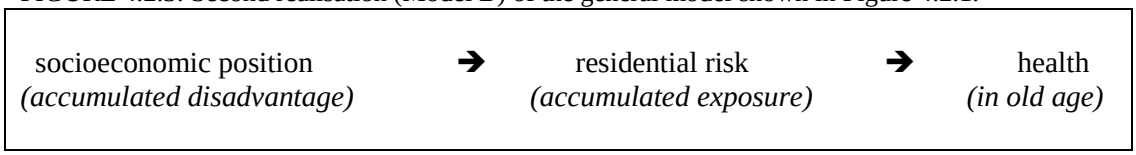
FIGURE 4.2.2: First realisation (Model A) of the general model shown in Figure 4.2.1.



In this model, the combined effect of multiple associations at specific points in time is hypothesised as exerting an influence on the person's state of health in old age. Model A may thus be thought of as incorporating elements of both critical period and accumulation of risk explanations of health inequalities. This in itself presents substantial challenges in analysis: the difficulties involved in 'disentangling' the respective effects of these two mechanisms have been illustrated (Hallqvist *et al.*, 2004).

A second model assessed in this study (Model B) is markedly different, in that it eschews completely the element of time-dependency which is central to Model A, and postulates instead relationships which are based purely on simple accumulations of risk (Figure 4.2.3).

FIGURE 4.2.3: Second realisation (Model B) of the general model shown in Figure 4.2.1.

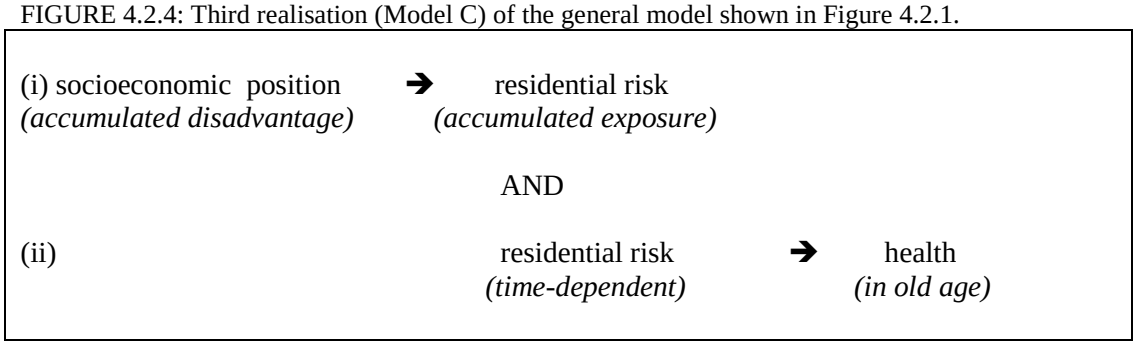


Model B replaces the set of time-dependent relationships featured in Model A with more general postulates. First, the individual's accumulated exposure to adverse residential conditions over the lifecourse is held to be significantly associated with her / his accumulated exposure to social disadvantage (assessed by the number of occasions, or number of years, during which s/he was considered to be subjected to such disadvantage). Second, accumulated exposure to housing hazards is assumed to exert a significant influence on health in old age. The concept of critical period effects is completely absent from Model B: it is the individual's total exposure to social disadvantage and to poor housing conditions which is of interest, irrespective of where in the lifecourse the exposure was experienced. Loosely speaking, Model A recognises patterns or trajectories of exposure, while Model B

allows consideration only of accumulations of exposure. Although conceptually more limited than Model A, Model B offers the opportunity to investigate associations (such as those between accumulated lifetime social disadvantage and health) which have not to date been exhaustively studied: "Many studies have demonstrated the graded association between socioeconomic position and health. Few of these studies have examined the *cumulative* effect of socioeconomic position throughout the lifecourse." [emphasis added] (Heslop *et al*. 2001, p. 477).

A final model assessed during the study (Model C) is shown in Figure 4.2.4. This is in essence a hybrid of Models A and B, being based on the hypothesis that accumulated social disadvantage is associated with health in old age; that accumulated disadvantage is related to accumulated exposure to poor housing; and that the individual's trajectory of exposure to housing hazards (reflecting the duration and temporal location of exposure) is linked to old-age health. The leftmost association in Model C is the same as the leftmost relationship in Model B, while the rightmost association in Model C is identical to the corresponding relation in Model A (though shown in more concise form in Figure 4.2.4).

FIGURE 4.2.4: Third realisation (Model C) of the general model shown in Figure 4.2.1.



While the models proposed above are concerned with associations among three factors (socioeconomic position, housing conditions and health), it was recognised that, as in most epidemiologic studies, estimation of these relationships was vulnerable to bias resulting from confounding. In particular, a large number of factors might reasonably be hypothesised to exert a confounding influence on the overall relationship between SEP and health. It was not realistic to attempt to investigate the influence of all such factors; rather, it was decided that the study should take account only of the possible confounding effect of (a) occupational risk (that is, exposure to work-related hazards such as fumes and physically demanding labour) and (b) smoking. Restriction of possible confounding effects to these two factors was a pragmatic decision, reflecting the information which was actually available in the data

source chosen for the study (see *Methods*, Chapter 5). Consideration of occupational exposures as a potential confounding influence reflected the plausible hypothesis that certain work-related hazards (notably exposure to fumes and other noxious substances) might exert a detrimental effect on respiratory health, which is one of the main areas of interest in the study[17]. The treatment of smoking as a possible confounder was justified by the well established role of tobacco smoke in the aetiology of both respiratory and cardiovascular disease.

Following the decision to include occupational exposures as potential confounding influences, execution of the study required the availability of information covering a total of five distinct factors: socioeconomic position, housing conditions, cardio-respiratory health (measured in later life), exposure to work-related hazards and smoking. The selection of a data source with the necessary attributes is described in Chapter 5.

***

Having outlined the hypothesis which the study was designed to assess, a formal statement of the research questions addressed by the study is now given.

4.3 Statement of research questions

In the previous section, three questions motivated by the conceptual model which underlies the study were identified. These, expressed in a more concise and formal form, represent the research questions which were treated by the study (see Box 4.3.1):-

BOX 4.3.1: Statement of research questions

1) How does exposure to social disadvantage, and to adverse residential conditions, vary over the adult lifecourse?
2) Do specific patterns of exposure to disadvantage and / or poor housing conditions predict cardio-respiratory health in old age?
3) How may detailed time-related patterns of exposure to the above risk factors be represented in epidemiological analysis of lifecourse influences on old-age health?

[17] While not wishing to anticipate introduction of the pre-existing data source which was used in the study (see Chapter 5), the documentation accompanying the dataset states, in connection with the recording of occupational fume and dust exposures, that "known respiratory hazards were the main focus." (*Users Guide to the Dataset*, p. 3)

This definitive statement of the study's objectives concludes the presentation of introductory material. Discussion now progresses to the methods which were developed to answer these questions.

CHAPTER 5: METHODS (I) - DATA SOURCE AND IDENTIFICATION OF VARIABLES

5.1 Data requirements

Assessing the three conceptual models introduced in Section 4.2, and addressing the research questions of Section 4.3, required the availability of a data source with three essential attributes. First, a detailed representation of the individual's socioeconomic position over an extended period of time was required. Second, the data needed to allow the subject's exposure to residential hazards to be determined to the same level of temporal detail as her/his trajectory of socioeconomic position. That is, if SEP was available at the level of the individual year, then hazard exposure measures were also required on a year-on-year basis. Finally, health measures which capture aspects of cardio-respiratory health in old age were demanded.

Clearly, it was not realistic to acquire such data on a prospective basis: several tens of years would be needed to collect the necessary information. Consequently, conduct of the study was entirely dependent on the availability of a pre-existing data source with the required properties. Literature was reviewed in an attempt to identify a suitable data source, the *Cohort Profile* series in the *International Journal of Epidemiology* being particularly useful for this purpose. It became apparent that data possessing the first two essential attributes (detailed information on SEP and on residential hazard exposure over a period of many years) are extremely rare; indeed, almost unknown. Consideration was given at an early stage to the possibility of using one of the major British cohort studies (such as the National Child Development Study, the 1970 British Cohort Study or the British Household Panel Survey). However, while these major data collection initiatives offer the considerable advantages of careful design, rigorous execution and substantial sample size, they could not satisfy the three requirements identified above. It was ultimately decided that the only pre-existing publicly-available data source which appeared to meet fully the requirements of the study was the Boyd Orr lifegrid sub-sample (Blane *et al.*, 1999; Blane, 2005). An outline of this resource is given in the next section.

5.2 The Boyd Orr lifegrid sub-sample: provenance and key features

A detailed account of this dataset (hereafter referred to as the 'BLS') is given in the references cited above. In essence, the BLS holds data obtained from a sample of 294 individuals who were originally included (as children) in a study of childhood diet and nutrition which was conducted by Sir John Boyd Orr in 1937-1939. A subset of these individuals was re-contacted in 1997-1998 (when the subjects were aged 63 to 78 years), and information on a wide range of their lifetime experiences elicited retrospectively. Areas covered included subjects' occupational histories (including information on job-derived social class, and exposure to work-related hazards) and residential experiences (including exposure to damp housing and air pollution). A range of health measures, captured in old age, was also collected. The technique used to elicit the retrospective data on occupation and housing was the lifegrid method (Blane, 1996). This seeks to guide the recollection of past events and experiences by "…cross-referencing the dates of any changes in the areas of interest, for example occupation and housing, against dates in the subject's personal life, such as marriage and death of mother, as well as against events in the external world, like coronations and wars." (Blane 1996, p. 753). The accuracy of data elicited via the lifegrid method has been examined in the specific context of traced participants in the original Boyd Orr survey (Berney & Blane, 1997). This investigation concluded that occupational and residential information (of central importance in the present study) was recalled via the method with useful accuracy.

When considering the BLS as a candidate data source, it was essential to confirm that it could provide the appropriate data elements needed to investigate the specific associations postulated in the study. It was also important to determine whether the sample of individuals featured in the BLS was characterised by extreme or unusual features which might seriously limit the generalisability of any inferences drawn. With regard to the first of these points, examination of the 466 variables in the dataset (both by direct scrutiny and by perusal of the *Users Guide to the Dataset*) confirmed that the required data elements were either directly available, or could be constructed via appropriate manipulations. An outline of the variables selected for use in the study is given in Section 5.3. The second point essentially reduces to the question of whether the BLS subjects were broadly representative of the population cohort from which the sample was drawn. This important aspect of the data has been investigated in some depth by Blane and colleagues, who concluded "those interviewed are shown to be representative of the British population socio-demographically… …and

physically." (Blane *et al.* 1999, p. 117). Sources of possible bias were identified in the data, but these were judged not to limit its usefulness:-

> "Bias is conservative because the most disadvantaged were disproportionately affected by loss to follow-up through death and because non-responders to interview were more disadvantaged as children than the interviewees. Representativeness and conservative bias, it is argued, justify the use of these data for investigating life course influences on health in early old age." (Blane *et al.* 1999, p. 117)

On the basis of the foregoing, it was decided that the BLS was an appropriate data source for use in the study. In adopting the dataset for use, analytical limitations - specifically, a lack of statistical power - imposed by the small available sample size (294 cases) were recognised and accepted as unavoidable.

Although the dataset contains both socioeconomic and residential information covering the subject's entire life (i.e. from birth to the point at which s/he was interviewed in old age), it was decided to restrict the age range examined in the study. This decision was motivated by the following argument. The representations of socioeconomic position which are available in the dataset are occupation-based, information on occupational class being available for each job held by the subject over her / his working life (see Section 5.3.1 below). Because the respondent's SEP was effectively to be defined by occupation, it was considered appropriate to constrain the age range examined to that part of the lifecourse throughout which s/he might potentially have been in employment (and thus be assigned to an occupational class). It was anticipated that enforcing this restriction would minimise (though not eliminate) problems associated with a major limitation of occupational class, namely its inability to accommodate those who are not in employment (see Section 1.3.2 in the *Introduction*). After consideration, it was decided that the period between 15 and 60 years of age could reasonably be regarded as corresponding to adult working life, and that this would be fixed as the period over which operation of the model shown in Figure 4.2.1 would be investigated. In reaching this decision, it was recognised that this range might have been extended at both extremes. Information on socioeconomic position in childhood was available (in the form of parental occupational class), and for those subjects who were retired at the point of data collection[18], the last occupation held could have been used to represent their SEP over the time period subsequent to retirement. However, it was felt that such rather imprecise imputation of SEP at the ends of the age range might dilute the very precise

[18] As stated earlier, respondents were aged between 63 and 78 years when health outcome information was collected.

occupation-based information which was available for the period during which the subject was employed. Moreover, it was analytically convenient to ensure that the length of time over which operation of the model would be assessed was the same for all subjects. The two goals of (a) restricting analysis to that portion of the lifecourse for which the most precise SEP information was available, and (b) standardising on a uniform age range for all subjects, could be achieved by effectively left-censoring the data at age 15, and right-censoring at age 60. Because the latter is the official state retirement age for women in the UK, this range approximately represented that portion of the lifecourse over which all subjects might potentially have been employed[19].

5.3 Basic data elements used in the study

The conceptual model described in Section 4.2 involves associations among three factors: SEP, housing conditions, and health. In addition, two further factors (experience of work-related hazards, and smoking status) were identified as potential confounders of the overall relationship postulated to exist between SEP and health. The variables used to portray each of these five factors in the study are now described in turn. For all factors, this section describes only those variables which were either directly available in the dataset, or could be derived via trivial manipulations. These are referred to as the *basic data elements*. The operations which were applied to these variables to construct the data structures used in the main analyses (generally consisting of either sequences or accumulated totals of exposure) are discussed at a later stage. The more complex measures resulting from these processes will be referred to as the *analytical data elements*.

5.3.1 Socioeconomic position

The BLS contains a number of variables describing the subject's socioeconomic position, based on the Registrar-General's occupational social class scheme (referred to hereafter as 'SC', following the convention introduced in the *Introduction*). The characteristics and limitations of this classification have been discussed earlier (see Section 1.3), and are not considered further here. For the purposes of the study, the descriptors of SEP used were:-

[19] Reflecting social conditions prevailing at the time, a high proportion of subjects actually entered employment before the age of 15 (see Section 6.1.2).

a) The occupational class of the subject's father interpreted according to the SC scheme. The permitted values for this variable are 'I', 'II', 'III (non-manual)', 'III (manual)', 'IV', 'V', 'Armed forces' and 'No paid work / inadequately described'.

b) The occupational class associated with the subject's first (i.e. chronologically earliest) job. The permitted values for this variable are as for father's occupational class, except that the category 'No paid work / inadequately described' (which is available for the latter) is omitted.

c) A further family of variables (defined as for [b] above) representing the occupational class associated with the subject's *N*th job in chronological sequence. The maximum number of jobs (and hence job-related occupational classes) recorded for any subject in the dataset is 11. The start and end dates associated with each job are recorded (as calendar years, e.g. '1944'), and multiple non-contiguous periods of employment are permitted i.e. the person may pursue job *A*, move on to job *B*, then subsequently return to job *A* at a later date. However, the occupational class associated with each job obviously remains constant across such multiple periods of employment in that job.

d) The main occupational class of the subject's first spouse. Permitted values for this variable are as for father's occupational class, with two differences. First, the category 'Armed forces' in the father's scheme is altered to 'Armed forces / National Service'. Second, an additional category 'Housewife' is added.

e) The main occupational class (defined as for [d] above) of the subject's second spouse.

5.3.2 Housing conditions: dampness and air pollution

The representation in the dataset of exposure to residential hazards is closely related to the method used to record the respondent's housing history. In essence, the dataset records the start and end dates (as calendar years) of the person's period of residence in each of the dwellings s/he lived in. The maximum number of dwellings recorded is 13. Multiple non-contiguous periods of residence in a single dwelling are permitted: the person may live in

home *A*, move to home *B*, then return to home *A*. For each dwelling, the following data elements are recorded in the dataset:-

a) A pair of variables holding the respective numbers of years during which the individual was 'probably' or 'possibly' exposed to dampness in that dwelling. The definition of these probability states is complex, and best expressed by direct quotation from the *Users Guide to the Dataset*:-

> "Residential damp was based on subjects' recall of the presence of black mould or other signs of damp. A 'damp probable' score was assigned if a subject recalled damp as being present in any of the main living areas of a house such as a living room or bedroom. Damp only in hallways or bathrooms resulted in a 'damp possible' score. The score only refers to the years in which damp was present. Thus, in some cases, the damp score may be less than the total number of years residence at one address. This would be due to the damp being eradicated during the residency." (*Users Guide to the Dataset*, p. 3)

b) A further pair of related variables containing the numbers of years over which the respondent experienced air pollution (probable or possible) at the dwelling. As before, the probable / possible distinction is most effectively described by direct quotation from the *Users Guide to the Dataset*:-

> "Air pollution measures were based on the level of urbanisation in the area of residence and proximity to industry or main roads. Subjects who described a pre-1960 residence as being in an urban area which had industry with emissions within a mile were assigned an 'air pollution probable' score for the number of years in which they lived at that residence. If a subject described a residence as being in an urban area with no industry with emissions within a mile then they were assigned an 'air pollution possible' score for the number of years in which they had lived at that residence. Following the Clean Air Acts of the 1950s, the main source of air pollution in the U.K. began to change from industry to road traffic. Consequently, for residences after 1960 subjects were only assigned an 'air pollution probable' score if they lived in an urban area and were within 150 metres of an A-road. An 'air pollution possible' score was assigned if they lived in an urban area, but were not within 150 metres of an A-road." (*Users Guide to the Dataset*, p.3)

5.3.3 Health

The health outcome variables available in the dataset are as follows:-

a) Whether the subject currently (i.e. at the time of interview, in early old age) suffers from any long-standing illness, disability or infirmity. Responses are restricted to YES and NO.

b) Whether such long-standing illness limits the subject's activities. Responses are again limited to YES and NO.

c) The nature of the long-standing illness(es) experienced. Details are recorded for up to four such conditions, the permitted values for each of these four variables being: heart disease; joint disease; lung disease; stroke; diabetes; abdominal hernia; thyroid disease; duodenal ulcer; pneumonia; high blood pressure; cancer; and any other illness. For the purposes of analysis, these original variables were re-expressed as a series of twelve binary indicators representing the presence (or absence) of heart disease, joint disease etc.[20]

d) Whether the subject is currently taking anti-hypertensive medication (YES or NO).

e) Whether the subject is currently taking bronchodilator medication (YES or NO).

f) Whether the subject is currently taking any prescribed medication (YES or NO).

g) The subject's systolic blood pressure (average of two readings, taken one minute apart).

h) The subject's diastolic blood pressure (again, average of two readings separated by one minute).

i) The highest value of the subject's forced expiratory volume (FEV) in one second, attained during at least three attempts.

j) The best value of the subject's forced vital capacity, again obtained over at least three repetitions of the manoeuvre. For analytical purposes, this variable was

[20] It was found that no subjects reported experience of pneumonia. Therefore, this variable was not used in the study and the number of variables indicating long-standing illnesses was reduced from 12 to 11.

combined with (i) to create a measure of FEV standardised for lung size, the relation being

$$FEV_1\% = (FEV_1 / FVC) \text{ X } 100. \text{ (Cotes } \textit{et al.}\text{, 2006)}$$

As explained in Section 4.2, the study concentrated on those measures which express aspects of cardio-respiratory health. These measures are: the presence of heart disease, lung disease, stroke and high blood pressure (four individual indicators; item [c] above); use of anti-hypertensive medication (item [d]); use of bronchodilator medication (item[e]); systolic blood pressure (item [g]); diastolic blood pressure (item [h]); and standardised FEV_1 (item [j]). Summary statistics for these quantities are given in Chapter 12 in the *Results* section of this thesis.

5.3.4 Exposure to occupational hazards

Information on the subject's exposure to work-related hazards is closely related to the representation in the dataset of the respondent's personal work history (see items [b] and [c] in Section 5.3.1 above). For each of the *n* jobs held by the respondent, the following information is recorded:-

a) A pair of variables holding the respective numbers of years during which the person incurred 'probable' or 'possible' exposure to occupational fumes and / or dusts in that job. The definition of the two probability states is not mathematical but conceptual, and is most effectively conveyed by quoting from the *Users Guide to the Dataset*:-

> "Occupational fumes and dust scores were based upon the type and level of exposure that a subject described as being commonplace in each of their particular jobs. For obvious hazards, such as asbestos, a 'fumes and dusts probable' score would be assigned. However, if, for example, a subject had been in a job for four years, but during that time had only been exposed to the hazard occasionally, the score would be adjusted accordingly. This could mean either a lower number of years for a 'probable' score or the full number of years being reduced to a 'possible' score. Similarly, part-time work resulted in adjusted hazard exposure scores. All occupational hazard scores were adjusted to account for low levels of exposure in full-time jobs, part-time work and 'seasonal' exposures." (*Users Guide to the Dataset*, p. 3)

b) The number of years during which the respondent was engaged in physically arduous labour in the job. For this hazard, no probable / possible distinction was

observed (the work either was or was not physically demanding for the number of years recorded).

c) The number of years during which the subject experienced demand / control stress (lack of job autonomy) in the job. As with (b) above, no distinction was made between probable and possible exposure.

In addition to the above, the dataset contains three further variables which record the subject's total accumulated lifetime exposure (in years) to, respectively, fumes / dusts (probable and possible combined), arduous work and demand / control stress.

5.3.5 Smoking status

The subject's smoking status is represented in the dataset by a categorical variable defining the individual's smoking status as one of 'current smoker', 'ex smoker' and 'never smoked'. In analysis, this was reduced to a simple binary contrast: current or former smokers *vs.* 'never smokers'.

***

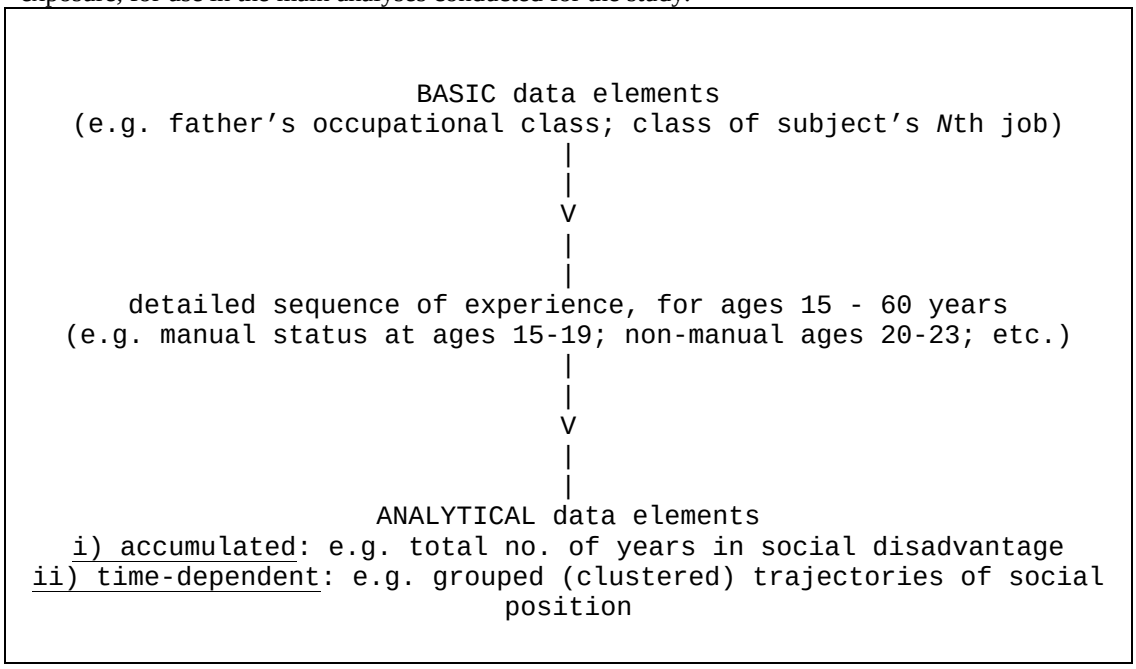
Having described the basic data elements selected for use, the process via which these variables were used to construct the analytical data elements (that is, the measures featured in the main analyses performed for the study) is now introduced.

5.4 Creating the analytical data elements: outline of process

In order to test the three conceptual models outlined in Section 4.2, the basic data elements described above were used to create more complex measures. While no modification of the health outcomes (Section 5.3.3) was required, the original plan for the study required that each of the other two main factors (SEP and exposure to residential risks) be realised in two forms. First, *accumulated measures* representing the subject's total exposure to each putative hazard (social disadvantage, dampness and air pollution) over the period examined were demanded. Second, measures representing subjects' *trajectories of experience* of SEP, and of residential risk exposure, were required. The remainder of the present section provides an overview of how these measures were derived.

Generating the analytical data elements (that is, accumulated measures and trajectories) involved a two-stage process. First, the basic data elements (described in the previous section) were manipulated to create interim data structures representing the factor of interest (e.g. social position) at each individual age point in the subject's life between the ages of 15 and 60[21]. Then, these detailed patterns or sequences of year-on-year experience were further refined to create both accumulated and time-dependent representations of the factor which were suitable for use in analysis. The process is illustrated in Figure 5.4.1. While the detailed sequences or trajectories of experience (the central element of Figure 5.4.1) were originally conceived as providing an interim stage in the creation of the final measures, it became apparent as the study progressed that these sequences did themselves incorporate information of direct relevance to the theme of the project. Consequently, the emphasis of the study changed somewhat over time: the detailed sequences came to be viewed as *outcomes* in their own right, as well as the product of an interim processing stage.

FIGURE 5.4.1: Outline of process adopted to create measures of social location and residential hazard exposure, for use in the main analyses conducted for the study.



Discussion continues by describing in detail the application of the process shown in outline form in Figure 5.4.1. Many of the considerations involved are common to both factors (i.e.

[21] The reasoning behind restricting the study to this age range has been presented at an earlier stage (see Section 5.2).

SEP and housing conditions), and the bulk of the explanation is provided in connection with the derivation of analytical data elements representing the former (Chapter 6 below).

CHAPTER 6: METHODS (II) - MEASURES OF SOCIAL POSITION

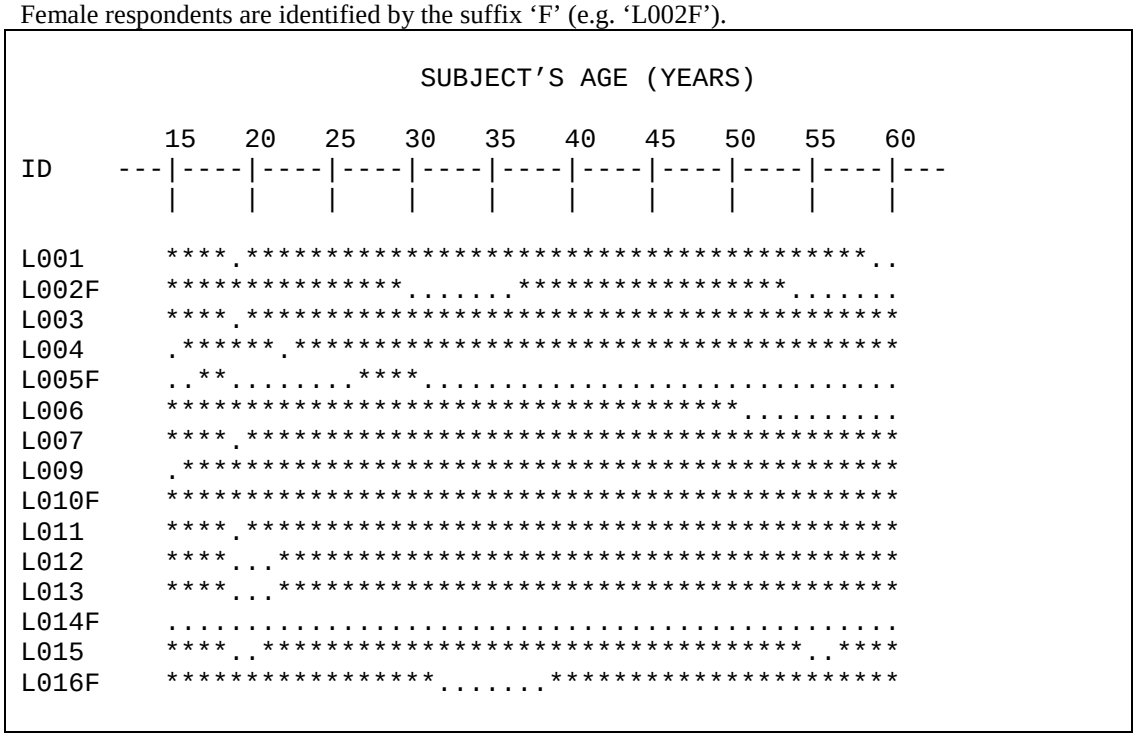
6.1 Constructing a detailed sequence of social position over time

6.1.1 Personal occupational class and periods of non-employment

As described in the previous chapter, construction of the main measures of social position was performed via an interim stage which involved creating a detailed sequence depicting the subject's SEP over time. Specifically, the requirement was to construct, from the variables listed in Section 5.3.1, a data structure holding an indication of the respondent's social position at each individual yearly point in the age range of interest (i.e. from 15 to 60; 46 datum points). Initial examination of the data indicated that an adequate representation of year-on-year social location could not reliably be created purely on the basis of the subject's own occupational history, because a high proportion of respondents were not recorded as being in continuous employment. Where an individual's employment record contained large gaps, a sequence of social position derived only from her / his job history would be badly fragmented. An example is provided by subject P051. This individual did not undertake her first job until the age of 45; therefore, a scheme based solely on job-derived occupational class would provide no information about a very large portion of her life, effectively excluding her from analysis. A fuller illustration of the limitations of personal occupational class as a descriptor of lifetime social position is provided by Figure 6.1.1 *(next page)*. This shows, for the first 15 subjects in the dataset, their employment status (that is, whether they were employed or not) at each year in the age range from 15 to 60. It is clear that the completeness of personal employment histories varies widely: for some subjects (e.g. L007), periods of non-employment are minimal, while for subject L014F no employment details at all are recorded.

The treatment of non-employment is a major challenge in occupation-based systems of social classification. The problem is highlighted by Galobardes *et al*. (2006a), who identify some common responses. One approach is to assign previous occupation to non-employed periods, thus assuming that a kind of 'social inertia' applies: the person retains her / his presumed level in the social hierarchy, even though personal circumstances have changed. This solution was not considered appropriate when designing the present study, as it

FIGURE 6.1.1: Representation of subjects' employment status across the age range from 15 to 60 years (restricted to first 15 respondents in dataset). Explanation of symbols appears beneath figure. Female respondents are identified by the suffix 'F' (e.g. 'L002F').



NOTE: Symbols used are '*' = subject employed at age *Y* years, '.' = subject non-employed at age *Y* years.

would have thwarted one of the major objectives: that of examining *in detail* individual trajectories of social experience. The loss of detail resulting from the wholesale replacement of periods of non-employment with the prior occupational class would have led to under-estimation of the number of state transitions (that is, shifts from occupational class *A* at age *Y* years to class *B* at age *Y*+1 years) present in subjects' social trajectories. For example, a progression of states such as

SC class III [ages 15 to 20 years] ➔

non-employed [ages 21 to 30 years] ➔

SC class III [ages 31 to 40 years]

would effectively be treated as a single continuous 'run' of 26 years' duration within class III, and the number of social transitions experienced by the respondent would be under-recorded by two. The number and timing of social transition events developed into one of the main interests of the study, and adoption of the 'previous occupation' approach would have resulted in these events being under-recorded in situations such as that illustrated above. Because of this, it was decided that non-employment should be retained as a distinct social state, rather than being obscured via the blanket imputation of the individual's prior

occupation. However, this decision generated certain challenges relating to how such periods of explicit non-employment should be interpreted.

While it would be possible to regard periods unaccounted for by the subject's personal employment history as simply indicating a state of conventional 'unemployment' (i.e. the inability to secure or perform paid work) at these times, determination of social position on this basis would be highly questionable. There are many reasons, both voluntary and involuntary, for an individual's non-participation in the labour market. While some of these (such as being unable to work through ill-health, or through simply not being able to find employment) are potentially associated with 'low' social status and financial adversity, others are not. Most notably, many women absent themselves from (or restrict their participation in) the labour market to satisfy the requirements of rearing children or running the household: "…wives' domestic responsibilities are crucial in severely limiting their employment opportunities." (Brown *et al*. 1985, p. 112). In particular, it may be argued that many women in this study are likely to have followed what came to be called in the 1960s the 'new conventional' pattern of employment: "…full engagement in the labour market or as students before acquiring 'housewife' status, generally a break during the years of starting a family, then resumption, but mainly part-time." (Fogarty 1985, p. 191). In such cases, the low social status and financial adversity associated with unemployment cannot be assumed automatically to apply during those periods when a woman (especially if married) was not actively employed. For this reason, it was decided that the determination of a subject's social position throughout life could not reliably be based solely on the occupational class(es) associated with her / his job(s), accompanied by an automatic presumption of low ('unemployed') social status at all times when no employment was recorded. Recognising this, it was decided that while the subject's personal occupational class would form the starting point for an initial characterisation of life-time social position, other data elements would be used where possible to generate a more complete representation. In other words, where the subject was employed at age *Y* years, her / his social position would be defined by the occupational class (I, II etc.) associated with the job held at that age. However, for ages where the subject was not employed, social position at those points would be determined by other means where feasible.

6.1.2 Parental occupational class and personal class in early working life

In view of the above limitations of personal (job-derived) class, it was decided that data representing parental (father's) class (see Section 5.3.1) would be exploited to provide additional information relating to the individual's social position in the early part of the period examined. While valid values of father's class were available for 290 of 294 cases, the intended use of this variable introduced the question of how far (in time) the influence of parental occupational class should be considered to extend. This issue arose because preliminary exploration of the dataset revealed that almost half of the subjects (143 of 290 cases [4 missing values]) reported taking up their first occupations at the age of 14. While this reflects social norms prevailing at the time (as demonstrated by the Factory Act of 1937, which stipulated that young people under 16 should work *no more than* 44 hours per week [Stevenson, 1990]), it is debateable whether at such a young age the social class associated with employment can be considered sufficiently influential to over-ride the person's parental social class. Thus, a challenging question arose: where the subject was recorded as being in work at the age of (say) 14 or 15, should s/he be considered to hold her / his personal occupational class (i.e. that associated with the job), or should the parental occupational class be held to apply? This point is of some importance, because an appreciable degree of discordance was evident in the dataset between parental social class and the class associated with the subject's first occupation. Of the 143 respondents who reported entering employment at the age of 14, information on both parental (father's) social class and the class associated with their first employment was available for 140. Among these cases, subgroups exhibiting a lack of concordance included 16 subjects reporting their fathers' social class as III manual, and their own class (at age 14) as III non manual. Similarly, a further group of 24 respondents indicated that while their fathers' social class was III manual, the class attributed to their own initial employment was IV. Thus, apparent instances of both upward and downward social mobility at the transition point of entering the labour market are present, but only if the individual may be assumed immediately to 'adopt' the social class associated with her / his job at an age as young as 14 years.

Arguments might be advanced in favour of permitting either the parental or personal occupational class to be considered dominant in such cases. However, it was decided that for the purposes of deriving detailed sequences of social location, the parental (father's) class would be deemed to apply until the respondent reached the age of 18 (i.e. up to and including 17), irrespective of whether the subject was or was not employed for a period prior to

attaining that age. Precedents for the imputation of father's occupational class to the individual in age ranges comparable to that used here are found in the literature. For example, in a study conducted by Ljung & Hallqvist which investigated associations between lifetime socioeconomic adversity and myocardial infarction, father's occupational class was deemed to define the individual's personal social location up to the age of 16 years (Ljung & Hallqvist, 2006).

6.1.3 Occupational class and married women

The decision to impute social position at early ages on the basis of parental class raised the question of whether the subject's social location should be assigned more generally on a 'highest in household' basis. For example, where an individual's own (job-derived) occupational class varied over the lifecourse between SC classes II and III, but the person's spouse was deemed to be placed in class I, it might be argued that the higher class was more representative of the individual's true location in the social hierarchy. The argument has particular appeal in the case of married women respondents, on the grounds that the social status of a male head-of-household (particularly at the period under consideration) would most accurately represent the 'true' social location of his wife. This view reflects the widely-applied (though sometimes acrimoniously-contested) approach to representing women in social stratification theory which has been summarised thus:-

> "Since women, it is argued, are peripheral to the occupational structure because of both their intermittent employment patterns and their primary responsibility for familial duties, their social class or prestige position is determined by the occupation of the bread-winner or the male head of household." (Hayes & Miller 1993, p. 654)

Although this view of women's social position has been the subject of heated debate (Sorensen, 1994), the imputation to women (and in particular to married women) of their partners' socioeconomic position is widespread in social science research (Krieger *et al.*, 2001; Beebe-Dimmer *et al*., 2004; Galobardes *et al*., 2006a). A *vignette* illustrating a plausible use of the husband's occupational class in preference to a woman's own under a specific set of circumstances is given by Bartley: "A woman married to a doctor who works as his secretary has a different 'standing in the community' from one who is simply a secretary and no more than that." (Bartley 2004, p. 144). A similar point is made by Brunner and colleagues when reporting a study investigating links between past and present socioeconomic circumstances and cardiovascular risk: "It should be borne in mind that employment grade may not be an accurate measure of current household circumstances, for

example, among women in clerical employment." (Brunner *et al*. 1999, p. 760). While acknowledging the validity of these considerations, the approach of using spouse's class in the present study was rejected because it would essentially discard the main advantage of the Boyd Orr lifegrid dataset, namely the extremely detailed information on occupational class which is available across the lifecourse. For example, suppose it were decided that a woman's social location should be determined primarily on the basis of her husband's occupational class. The only information available in the dataset on spouse's class is (a) occupational class at marriage; (b) 'main' occupational class (that held throughout the greater part of working life); and (c) class at retirement. If the husband's class were used in preference to the (female) subject's own job-derived class, considerable detail would potentially be lost: the husband's blanket 'main' class might in fact mask considerable fluctuations in social location across the lifecourse. If the husband's class history were available to the same level of detail as the subject's, there would be an argument for using the former under certain circumstances for married females. However, since this is not so, the approach ultimately adopted was to use the husband's general measure of main class only as a secondary indicator of social position, to complete 'gaps' in the woman's own, more detailed, job-derived class history. That is, where the social location of a married woman at age *Y* years could not be determined because no employment was recorded for that age, her husband's main occupational class was considered to apply.

While there is extensive precedence for the imputation to women of their partners' occupational class (Krieger *et al.*, 2001; Beebe-Dimmer *et al*., 2004; Galobardes *et al*., 2006a), there is no corresponding accepted tradition of the reverse process (i.e. applying to men the occupational class of their wives). Consequently, it was not considered appropriate to complete gaps in the occupational histories of male respondents by imputing to them the main class of their wives.

6.1.4 Simplifying the sequences: dichotomising occupational class

Reflecting the decisions described above, detailed sequences representing the individual's social position (according to the SC scheme) at each yearly age point in the range from 15 to 60 were constructed on the basis of:-

a) the subject's own (job-derived) occupational class at age *Y* years

b) the occupational class of the respondent's father (considered to apply over the age range from 15 to 17, over-riding any personal class information at these ages)
c) for married women only, the main occupational class of their husbands (applied only when no personal class was recorded at age *Y* years)

A sample of the results from this process is presented in Figure 6.1.2, which shows sequences of socioeconomic position for the 15 subjects featured earlier in Figure 6.1.1. Comparison of the two figures demonstrates how the imputation of parental class and spouse's main class has increased the completeness of the sequences (notably for subjects L005F and L014F).

FIGURE 6.1.2: Representation of subjects' socioeconomic position across the age range from 15 to 60 years, after imputation of parental class and spousal class where appropriate (restricted to first 15 respondents in dataset). Explanation of symbols appears beneath figure.

```
                          SUBJECT'S AGE (YEARS)

          15   20   25   30   35   40   45   50   55   60
ID    ---|----|----|----|----|----|----|----|----|----|---
          |    |    |    |    |    |    |    |    |    |

L001      4444.444444466666666666666666666666666666..
L002F     444444444444444222222255555555555555552222222
L003      4444.444444444444444444444444444444444444444
L004      4443333.333333333333331111111111111111111115
L005F     5555........333344444444444444444444444444444
L006      444555555555554444444444444444444.........
L007      4445.444444444444444444444444444444444444444
L009      333444444444444444444444444444443333333333444
L010F     444555555555555555555555555555544444445555555
L011      6666.666664444444444422222222222222222444444
L012      5554...444444666655555555555555555444444444444
L013      6664...444444444444444444444444444444444444444
L014F     222.555555555555555555555555555555555555555555
L015      2222..444443334444444444222244433333333..2223
L016F     222333333333333331111111555555555555555555555
```

NOTE: Symbols used are '1' = SC Class I, '2' = Class II, '3' = Class IIINM, '4' = Class IIIM, '5' = Class IV, '6' = Class V, '.' = non-employed

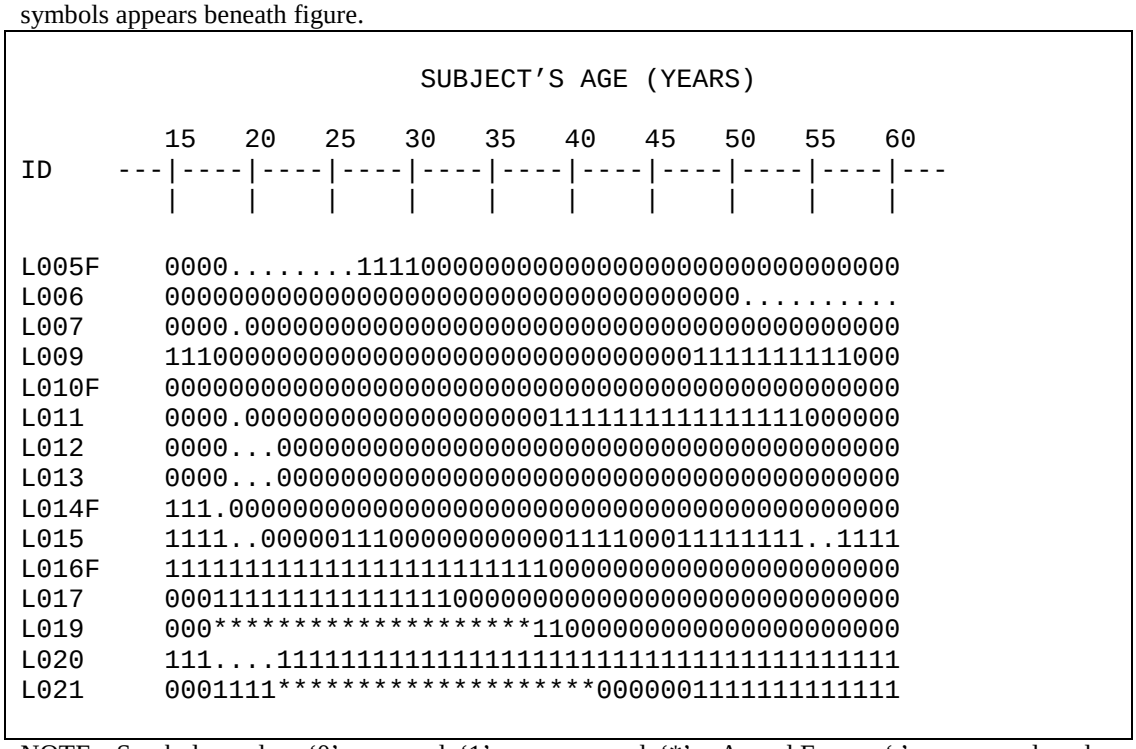
In preparation for the creation of the variables required in the study's main analyses, the sequences derived via the above process were subject to one further operation. The six classes within the SC scheme were dichotomised into a simple contrast between 'manual' (classes IIIM, IV and V) and 'non-manual' (classes I, II and IIINM). This approach is common in research into health inequalities (e.g. Davey Smith *et al.*, 1997; Heslop *et al.*, 2001; Mishra *et al.*, 2009; Watt *et al.*, 2009) and enjoys general acceptance (Galobardes *et*

al., 2006b). On completion of this step, the subject's socioeconomic position at each age point was defined as one of:-

- manual
- non-manual
- engaged in Armed Forces service
- non-employed

A sample of the resulting sequences is shown in Figure 6.1.3, which intentionally features a different subset of cases from that shown in Figures 6.1.1 and 6.1.2. The selection of subjects for Figure 6.1.3 was designed to ensure some representation of the relatively rare 'Armed Forces' state. Properties of the sequences created are presented in the *Results* section (Chapter 14).

FIGURE 6.1.3: Final representation of subjects' socioeconomic position across the age range from 15 to 60 years, after dichotomisation of SC classes (restricted to 15 selected respondents). Explanation of symbols appears beneath figure.



NOTE: Symbols used are '0' = manual, '1' = non-manual, '*' = Armed Forces, '.' = non-employed

6.2 Deriving a measure of accumulated disadvantage

6.2.1 Representations of accumulated disadvantage in the literature

As stated earlier, the detailed sequences of social position were originally envisaged as an interim step in the process of creating the variables needed to assess the conceptual model postulated in the study. One of these required measures was a quantity representing the subject's level of accumulated exposure to social disadvantage. A wide range of precedents exists for the construction of such measures in research into health inequalities. One example is a study by Davey Smith and colleagues which investigated the influence of lifetime socioeconomic position on mortality from various causes (Davey Smith *et al*., 1997). In this, the social class of subjects (defined as manual or non-manual) was established at three points in life: from the social class of their fathers, their own class on entering employment, and the social class associated with the job held at the time screening for the study was conducted. From this information, an indicator of cumulative socioeconomic disadvantage across the lifecourse was derived by summing the number of points at which the subject's class location was manual (hypothesised as reflecting disadvantage) or non-manual. Thus, a subject might be identified as being (for example) non-manual at two points and manual at the third, or as manual at all three stages. The study then investigated relationships between this representation of cumulative lifetime disadvantage and a number of health outcomes.

A rather different approach to defining social disadvantage across the lifecourse was taken in an investigation of links between the lifetime accumulation of adverse socioeconomic position and the risk of myocardial infarction (Ljung & Hallqvist, 2006). In this study, year-by-year information on occupation-based SEP from birth to the onset of disease was available. The authors used these data to create a single index representing the proportion of life spent in adverse socioeconomic circumstances (defined as manual occupational class). This index was reduced for analysis to a five-way ordinal scheme ranging from 'never' (no time in adverse conditions) to 'always' (entire life in adverse circumstances). In passing, this study is noteworthy in another respect in the context of the present project, in that these authors explicitly draw attention to the fact that they used *individual* socioeconomic position (as distinct from household position) to define adversity. As discussed earlier (See Section 6.1.3), the issue of whether to use individual or 'highest-in-household' social location was also confronted by the present study. Further examples of studies which feature some

measure of accumulated socioeconomic disadvantage are those reported by Lynch *et al.* (1997), Power *et al.* (1999), Heslop *et al.* (2001), Pensola & Martikainen (2003), Naess *et al.* (2004a) and Singh-Manoux *et al.* (2004).

***

In the present study, deriving a measure of accumulated disadvantage from the detailed sequences of social position was not straightforward. Had representation of the individual's status at each age point been limited to either of the two possible states in a simple manual / non-manual dichotomy, the task would have been trivial, requiring only a simple calculation of the total number of years during which the person was deemed to be manual. However, the presence in the state space of two additional conditions (engaged in Armed Forces service, and non-employed) meant that a different approach was required. The treatment of these two problematical states is now discussed, beginning with Armed Forces service.

6.2.2 Simplifying the state space (i): treatment of Armed Forces service

Consideration was initially given to simply excluding from analysis those subjects for whom any period of military service was recorded, an approach for which ample precedent exists: "…a seventh [SC] category includes all people in the armed forces irrespective of their rank therein, which is generally excluded in health studies." (Galobardes *et al.* 2006b, p. 95). However, the policy of exclusion was considered inappropriate for the present study, since it would have inflicted an unacceptably high level of attrition upon what is, to begin with, a small dataset. Of the 294 subjects in the dataset, 41 recorded some degree of military affiliation (either personally or - for married women - via the imputation of husband's main occupational class). Excluding these individuals would entail the loss of 14% of the data. As an alternative, it was decided that for the purposes of calculating accumulated disadvantage, Armed Forces service should (with certain exceptions, described shortly) be treated as equivalent to manual status and consequently considered indicative of disadvantage. This forced equivalence of the manual and Armed forces states is justified on two grounds. First, it may reasonably be argued that manual and military occupations are both characterised by lack of job control; that is, the extent to which a worker enjoys meaningful autonomy in deciding the method and timing of execution of his work. In civilian life, studies have demonstrated that job control is socially patterned, with a lesser degree of control being experienced by those lower in the occupational hierarchy. For

example Marmot and colleagues, using data from the Whitehall II study, found that the proportions of men reporting low job control were 8.7% among those deemed to occupy high grade posts, 26.6% among those in intermediate grades, and 77.9% among men whose jobs were classed as low grade. (Marmot *et al.*, 1997b). A similar gradient was observed among female workers, and the authors concluded that "Low job control was closely linked to position in the employment hierarchy." (Marmot *et al*. 1997b, p. 237). A related conclusion was drawn by Bartley *et al*. when examining work control among female workers, using data from the Health Survey for England:-

> "...women in the service classes and the self-employed have more control at work than those in routine non-manual occupations while those in skilled manual jobs have more than those in less skilled manual jobs." (Bartley *et al*. 2000, p. 68)

In fact, the degree of socially-patterned variation in work control identified by Bartley *et al*. is striking. Basing their investigation on the Erikson-Goldthorpe class *schema*, the authors found that the odds ratios for experiencing low work control, relative to the Higher Professional class, were:-

Lower professional	-	2.74
Routine non-manual	-	13.33
Self-employed	-	1.09
Skilled manual	-	10.49
Non-skilled manual	-	19.30

(Bartley *et al*. 2000, p. 66 [data reproduced from Table 4.3])

Extrapolating these findings to the data used in the present study, it might reasonably be postulated that levels of job control will in general be rather lower among individuals in the manual state than those in the non-manual condition at any point in time. The concept that manual status is characterised by reduced levels of job control points the way to what is arguably an element of commonality between the manual and Armed Forces conditions. Clearly, 'job control' (or rather the *absence* of control) is a defining feature of military institutions. Such bodies by their very nature enforce a degree of authoritarianism, and a requirement to obey any order *without question*, which transcend those found in even the most strictly-run of civilian enterprises. While it is obviously unwise to draw too close a comparison between the often complete absence of job control which prevails in military settings, and the levels of autonomy which apply in lower-graded civilian occupational contexts, there is arguably a point of resemblance. For this reason, and while recognising the limitations of the comparison, it is argued that one justification for equating the Armed

Forces social state defined for this study with the manual state is that both are characterised by limited levels of job control. The justification perhaps becomes more persuasive if the definition of 'control' offered by Bartley *et al*. is adopted: "the degree of power which other people have over the conduct of an individual's working day" (Bartley *et al*. 2000, p. 59). Both the foreman or supervisor of low-grade employees, and the officer or NCO in command of soldiers, enjoy an appreciable degree of *power* over those subordinate to them.

A second area of resemblance between the manual and Armed Forces conditions, as defined in the present study, relates to the respective degrees to which each involves exposure to occupational hazards. At first sight, any such comparison is faintly ludicrous: obviously, military service is unique in that it involves systematic exposure to death and wounding intentionally inflicted at the hands of hostile elements. No civilian occupation is remotely comparable in this respect. However, leaving aside the element of exposure to *intentional* harm, it is in fact arguable that both manual occupations and military service are characterised by working conditions which are potentially injurious to health. In the case of the manual worker, the hazards involved include industrial accidents, strenuous or damaging manual labour and exposure to noxious substances (fumes, chemicals etc.), or to extremes of heat or cold. However, even when the specific risks of death or wounding associated with active military service are excluded, exposures incidental to army life can be extremely damaging to health. It is reasonable to assume that some of the subjects in the Boyd-Orr dataset will have served during the Second World War, and the health consequences of active service in that conflict could be devastating even where death or wounding were not involved:-

> "On July 17 [1945], after a particularly bitter series of signals, my demands for a medical examination of my brigade were granted. At nameless spots in the jungle, over the next three days, every man in the brigade was examined by medical boards consisting of two or three doctors. The strength of my four and a half battalions then totalled about 2,200 men. Those adjudged fit for any kind of action, in any theatre of operations, numbered 118…" (Masters 1961, p. 281 [describing the condition of his brigade after c. 110 days of operations in Burma]).

Here, then, is a second element of commonality between the manual and Armed Forces states: both potentially expose the individual to environmental conditions which may be damaging to health. As with the earlier comparison on the basis of limited job control, the resemblance cannot realistically be pressed too far. Consequently, it was not considered defensible to conduct all of the analytical work for this thesis under a blanket assumption of identity between the manual and Armed Forces states. Nonetheless, it is argued that the

degree of commonality between these two states is sufficient to justify their amalgamation for the purposes of deriving a measure of accumulated disadvantage.

Based on the above reasoning, the states featured in the detailed sequences of social position were modified so as to conflate the manual and Armed Forces conditions for the purposes of calculating accumulated disadvantage. For example, the original numbers of years observed in the four possible social states for subject P119 were:-

manual	-	3 years
non-manual	-	40 years
Armed Forces	-	3 years
non-employed	-	0 years

After forcing the period accounted for by Armed Forces services into equivalence with the manual state, the accumulated totals for this subject became:-

manual / Armed Forces	-	6 years
non-manual	-	40 years
non-employed	-	0 years

Clearly, this process involved the introduction of a degree of 'error', in that one relatively uncommon social state was coerced (for reasons of analytical convenience) into an artificial equivalence with a second, more commonly-observed, state. It is not unreasonable to assume that the level of uncertainty imposed by this process on analysis will relate to the magnitude of the values observed for the state which is forced into equivalence, and it is therefore instructive to gain some appreciation of the scale of Armed Forces service among subjects during the period of interest. Table 6.2.1 *(next page)* shows the distribution of these values, which suggest that respondents' recorded association with the Armed Forces divides naturally into three types:-

no service (253 subjects of 294; = 86.1%)
'short' service of 2 to 7 years' duration (32 of 294; = 10.9%)
'prolonged' service of 20 to 35 years' duration (9 of 294; = 3.1%)

Of the above, the group of greatest interest in connection with the forced equivalence between the Armed Forces and manual states is the nine subjects who recorded prolonged levels of the former. For these individuals, the mean duration of military service was 24.1 years (SD: 4.6 years). It was decided that the life experiences of these people were likely to

TABLE 6.2.1: Distribution of the total number of years during which subjects engaged in Armed Forces service between the ages of 15 and 60.

Armed Forces service (years)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	253	86.05	253	86.05
2	7	2.38	260	88.44
3	10	3.40	270	91.84
4	4	1.36	274	93.20
5	5	1.70	279	94.90
6	4	1.36	283	96.26
7	2	0.68	285	96.94
20	2	0.68	287	97.62
22	2	0.68	289	98.30
23	2	0.68	291	98.98
26	2	0.68	293	99.66
35	1	0.34	294	100.00

be markedly different from those of the bulk of the sample (i.e. subjects with limited military service, or none at all), and that for these nine respondents, forcing an identity between the Armed Forces and manual states would introduce an unacceptable degree of error. Therefore, these nine subjects were excluded from the calculation of accumulated disadvantage (i.e. assigned missing values).

6.2.3 Simplifying the state space (ii): treatment of non-employment

Treatment of the remaining 'problematical' state (that of non-employment) was more challenging. The distribution of numbers of years in non-employment for the $n = 285$ subjects retained after exclusion of those with extended Armed Forces service is shown in Table 6.2.2 *(next page)*. From the table, it is clear that (in contrast to the corresponding distribution of Armed Forces service values) there is no marked natural break between those reporting 'short' and 'prolonged' levels of non-employment. As a result, it is difficult to justify the exclusion from analysis of the latter, because any demarcation value above which subjects were deemed to experience prolonged non-employment would be largely arbitrary. Consequently an alternative approach was adopted, justified by the following argument.

It has previously been suggested (Section 6.1.1) that non-employment - as distinct from *unemployment* (loosely defined as the involuntary ability to obtain or perform paid work) - cannot automatically be held to indicate low social status and / or financial hardship. However, the process described in Section 6.1.3 effectively 'filtered out' many instances of

TABLE 6.2.2: Distribution of the total number of years during which subjects were non-employed between the ages of 15 and 60.

non-employment (years)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	157	55.09	157	55.09
1	39	13.68	196	68.77
2	18	6.32	214	75.09
3	16	5.61	230	80.70
4	13	4.56	243	85.26
5	13	4.56	256	89.82
6	9	3.16	265	92.98
7	4	1.40	269	94.39
8	6	2.11	275	96.49
9	4	1.40	279	97.89
10	2	0.70	281	98.60
12	1	0.35	282	98.95
18	1	0.35	283	99.30
21	1	0.35	284	99.65
29	1	0.35	285	100.00

non-employment by attributing to married women the main occupational class associated with their husbands. As a result, women who withdrew voluntarily from the labour market (e.g. for purposes of child-rearing) would, in the main, no longer be designated in the social sequences as non-employed. On this basis, it may plausibly be argued that the remaining instances of non-employment at age *Y* years are, in many cases, likely to be indicative of genuine unemployment as it is conventionally understood. If this is accepted, the three remaining states in the state space may reasonably be viewed as an ordered scheme, with non-manual representing the 'best' (least disadvantaged) status, manual / Armed Forces holding an intermediate position, and non-employed the 'poorest' (most disadvantaged). From this, it is arguably defensible to simplify the three-state scheme such as to contrast non-manual with the two remaining (disadvantaged) states in combination. Such an arrangement effectively compares those with no level of presumed social disadvantage against those who are considered to experience some degree of disadvantage. This was in fact the approach adopted. To illustrate, the numbers of years observed in the three states for subject P105 were:-

manual / Armed Forces	-	9 years
non-employed ('unemployed')	-	3 years
non-manual	-	34 years

Under the scheme just outlined, these values were modified to create a simple two-way representation:-

disadvantaged (manual / Armed Forces / non-employed)	-	12 years
not disadvantaged (non-manual)	-	34 years

On completion of this operation, a measure of accumulated disadvantage was constructed by summing the number of years (minimum zero; maximum 46) during which the respondent was subject to disadvantage (defined as above) over the age range of interest. The properties of this measure are summarised in Section 11.2 in the *Results* section.

6.3 Creating a measure of time-dependent social position

6.3.1 Analytical limitations of the detailed sequences of social position

In addition to the measure of accumulated disadvantage described above, the objectives of the study dictated that a representation of the individual's *trajectory* of SEP be available. The requirement was for a measure which would indicate the subject's social location at specific stages of the age range examined. For example, the respondent might hold manual class in the early part of the period (say, from 15 to 25 years), then enjoy non-manual status thereafter. This type of time-related variation is (obviously) completely obscured when social position is represented only by an accumulated measure, thus reducing a possibly complex and fluid experience to a single number.

Before proceeding to describe the derivation of trajectories of social location, it is helpful to highlight a point of terminology. In the discussion which follows, the term 'trajectory' is generally used (interchangeably with 'sequence') to describe a time-dependent pattern of experience (here, of social position; elsewhere in this thesis, of exposure to environmental hazards). This usage is slightly more general than that employed in the literature relating to sequence comparison in other disciplines, in which a trajectory is sometimes defined specifically as a pattern of change in a quantity which varies *continuously* in time, as distinct from one which is sampled at discrete intervals (e.g. years). This sense is expressed by Kruskal & Liberman:-

> "In speech processing, gas chromatography, bird song, and other potential applications of sequence comparison, the underlying objects of interest are basically continuous functions **a**(*t*), **b**(*t*), etc., of a continuous variable *t*, which is often time… …Thus each object of interest is a continuous *trajectory* or curve through feature space… …For

> practical manipulation, these trajectories are ordinarily converted into sequences by sampling the values of *t*." (Kruskal & Liberman 1999, p. 130)

The usage of 'trajectory' in this thesis is mirrored elsewhere in the social science literature (e.g. Adams *et al*., 2004; Hallqvist *et al.*, 2004; Naess *et al*., 2006), and adoption of this meaning is in no way controversial. However, this very minor point of terminology is highlighted here because it conveniently introduces the interesting question of whether social position may be regarded as a factor which can genuinely vary continuously over time. This in turn raises the issue of how frequently SEP needs to be sampled in order to provide an adequate characterisation of the individual's social position over an extended period of time. This topic is developed further in the *Discussion*.

The sequences of social position described in Section 6.1 are themselves highly detailed representations of subjects' social trajectories over the adult lifecourse, and the original intention was to use them directly, in their original form, in the main analyses. However, this proved not to be feasible due to the substantial (and unexpected) degree of diversity and variation observed in the sequence set. While it is not appropriate at this stage to present the characteristics of the sequences in detail (this being reserved for the *Results* [Chapter 14]), it was found that no fewer than 216 unique trajectories of social position were observed among the 294 cases in the dataset. The limitations imposed on analysis by this variety of experience are best illustrated by a simple example. Suppose it were desired to investigate the association between an individual's trajectory of social position and one of the binary health outcomes of interest, such as the experience of heart disease. In theory, this relationship could be assessed via a logistic regression model, using the sequence structure of Section 6.1 as the predictor (accompanied by whatever other covariates might be considered appropriate). However, because representing an explanatory factor with *n* levels in a regression model requires a group of *n*-1 binary indicators, no fewer than 215 indicators would be needed to represent the full set of trajectory types present in the data. This approach immediately presents two difficulties. First, such a model would not converge to completion due to sparseness in the data (leading to complete or quasi-complete separation of data points, and non-existence of a maximum likelihood estimator [Silvapulle, 1981; Albert & Anderson, 1984). However, even if this were not so, a second difficulty arises in that meaningful interpretation of 215 individual parameter estimates would be extremely challenging:

> "...the human mind, while it may be able to encompass say 10 numbers easily enough, finds 100 much more difficult, and will be quite defeated by 1000 unless some reducing process takes place." (McCullagh & Nelder 1989, p. 3)

To facilitate useful analysis of (for example) the influence of SEP trajectory on a measure of hazard exposure or a health outcome, it was clear that the large number of unique sequences of social location must be reduced to a smaller number of aggregated categories in which 'similar' sequences were grouped together in a common category. In other words, the 216 sequences found in the data must be condensed into n higher-level groups, where $n < 216$. The result of such a process would ideally constitute a *typology* of time-dependent social location, representing general patterns of social progression over time (e.g. 'upwardly mobile in later life') rather than detailed individual experiences. A very similar intention underlies a study of McVicar and Anyadike-Danes, which used methods closely related to those employed in the present study to "create a typology of youth transitions [from school to work]". (McVicar & Anyadike-Danes 2002, p. 319).

Discussion now proceeds to describe how condensation of the original sequence set into a notional typology of experience was achieved. However, before doing so it is useful to highlight that the root of the analytical problems identified above (i.e. the diversity and individuality of social experiences observed in the data) is regarded as a noteworthy finding in itself, which arguably carries interesting implications for research into health inequalities. This topic is treated at some length in the *Discussion* chapters of this thesis.

6.3.2 Grouping the sequences (i): classification as a research objective

Creating a typology of social location sufficiently compact for use in analysis is essentially a problem of *classification*; that is, of assigning objects (here, sequences of social position over time) to groups or classes in which each object in a group is similar to others in that class. The problem of classifying individuals or objects is not peculiar to research into health inequalities; a similar requirement arises in many other disciplines. Before considering examples, it will be helpful to establish some terminology. In the statistical literature, the term *classification* is (confusingly) used in two quite distinct senses (Venables & Ripley, 2002). These indicate, respectively, assigning objects to known groups which are defined *a priori*, and allocating objects to places within a structure of natural groups which is derived from the data themselves. A fuller definition of the first sense is given by Johnson & Wichern, who specify the goal of classification thus:-

> "Classification pertains to a *known* number of groups, and the operational objective is to assign new observations to one of these groups." (Johnson & Wichern 1992, p. 573)

The second sense is defined by Gordon, and the difference with the first sense explicitly emphasised:-

> "In classification… …the classes are not known at the start of the investigation: the number of classes, their defining characteristics and their constituent objects all require to be determined. The word 'classification' has also been used in a different sense, to refer to the assignment of objects to one of a set of already-defined classes. …each object is assumed to belong to one of a known number of classes, whose characteristics have been determined using a 'training set', and the aim is to identify the class to which the object should be assigned." (Gordon 1999, p.3)

The contrast between the two meanings is expressed more succinctly by Hardle & Simar:-

> "Given a data set containing measurements on individuals, in some cases we want to see if some natural groups or classes of individuals exist, and in other cases, we want to classify the individuals according to a set of existing groups." (Hardle & Simar 2007, p. 271)

In reporting the present study, classification is used exclusively in the sense of placing objects (in the present context, trajectories of SEP) into groups which are suggested by the data, not defined *a priori*. Classification in this sense is often termed *clustering* (Jain *et al.*, 1999; Fielding, 2006; Theodoridis & Koutroumbas, 2006) or *cluster analysis* (Johnson & Wichern, 1992; Timm, 2002; Hardle & Simar, 2007). Throughout this thesis, the terms classification, grouping and clustering / cluster analysis are used as synonyms to convey this concept.

The use of classification (in the sense just established) is common in social science research. This may be illustrated by considering some recent examples of published work based on this class of method. Sabbe *et al*. used K-means cluster analysis to identify subgroups of Flemish elementary schoolchildren based on their patterns of physical activity and eating habits. The resulting cluster scheme permitted subjects to be characterised as 'sporty healthy eaters', 'sedentary healthy eaters' and so on (Sabbe *et al*., 2008). The use of seven different clustering methods was explored by Kuehne *et al*. to investigate factors influencing irrigation practices among farmers and landowners in Australia's Namoi Valley (Kuehne *et al*., 2008). After selecting Ward's method as that which gave the most usefully interpretable results, these investigators identified a three-cluster solution which classified their subjects as

'providers', 'investors' or 'lifestylers'. A study of drug use and sexual HIV risk conducted by Schonnesson *et al*. located three clusters in a sample of HIV-seropositive African-American crack cocaine smokers (Schonnesson *et al*., 2008). Using the TwoStep dual-stage clustering algorithm provided in SPSS software, these authors classified their subjects as 'highest risk', 'consistent condom use' or 'inconsistent condom use'. As a final example, an investigation by Berry of community participation in a socioeconomically disadvantaged rural Australian community also used the TwoStep method to characterise subjects as 'busy working parents', 'social capital elites', 'excluded participators' or 'aging, participating less' (Berry, 2008).

Beyond the social sciences, the range of research disciplines and commercial applications in which classification plays a role is vast. Gordon (1999) cites examples of classification / cluster analysis being applied in fields as diverse as archaeology, plant ecology, taxonomy and (whimsically) the study of the properties of malt whisky[22]. Jain *et al*. (1999) instance the use of clustering in (*inter alia*) data mining, document retrieval, biology, psychiatry, geology and geography. Hardle & Simar (2007) describe applications of cluster analysis in marketing (to select test markets), insurance (to distinguish among classes of potential customers) and psychology (to identify types of personalities from questionnaire responses). Cluster analysis is also widely used in many areas of medical research; some examples are given by Shannon (2008). In the present study, the general role of classification is to group patterns of experience, with the goal of creating a concise typology of the factor of interest (in the present discussion, of social position over time) which is analytically more tractable than the individual trajectories from which the typology is constructed. However, the use of cluster analysis in an inferential (as distinct from an exploratory) role is potentially questionable, some authorities holding that the use of clustering techniques should be restricted to the latter. Because the present study makes extensive use of groups derived via classification in inferential analysis (e.g. determining whether health outcomes vary across clusters of lifetime SEP), the use of clustering in an inferential context must be justified. This topic is discussed in the next section, which also considers some more general criticism of classification as an analytical approach.

[22] In fact, the paper cited thus by Gordon (Legendre & Lapointe, 2004) is a highly technical discussion of testing for congruence among distance matrices; the whisky data are merely illustrative.

6.3.3 Grouping the sequences (ii): classification as a vehicle for statistical inference – limitations and criticisms

As with all statistical techniques, classification / clustering (in the sense defined above i.e. that of allocating subjects to groups which are derived from the data rather than defined *a priori*) is of value only when correctly applied. This section begins by considering criticism of classification (and of its misapplication in research) which has been expressed in the statistical literature, then proceeds to argue that the use of this family of methods in the current study is valid and defensible. A vigorous critique of the inappropriate use of classification was expressed as long ago as 1971 in a frequently-cited paper by Cormack:-

> "The summarization of large quantities of multivariate data by clusters, undefined *a priori*, is increasingly practiced, often irrelevantly and unjustifiably… …The availability of computer packages of classification techniques has led to the waste of more valuable scientific time than any other 'statistical' innovation (with the possible exception of multiple-regression techniques)." (Cormack 1971, p. 321)

This uncompromising sentiment received some support in the discussion which followed the paper's presentation[23], one commentator remarking "I know that all of us would agree that statistical classification is a curious mixture of sense and nonsense - although we might disagree about which methods should be assigned to which categories." (Sibson, quoted in Cormack [1971]). While Cormack's criticism of clustering includes many points of detail, one of his main objections is succinctly expressed in his conclusion:-

> "Clustering uses much time and effort. We want to cluster only if clusters exist... ...the ability of procedures to find non-existent clusters is well established." (Cormack 1971, pp. 345-346)

This property (that classification procedures will identify groups, even when these groups have no meaning) is recognised by other commentators e.g. Fielding:-

> "…before beginning any unsupervised clustering of unlabelled cases it is worth remembering that the algorithms will place cases into clusters, even when real clusters do not exist. This is because, in general, they do not 'give up' if there is no obvious structure. The rules continue to be applied until all cases are in clusters." (Fielding 2006, p.47)

[23] Cormack's paper was originally presented orally at a meeting of the Royal Statistical Society, and the published version included a full account of the discussion which ensued.

Here, then, is a first major limitation of cluster analysis: that groups will be located even where no clear structure is evident in the data. A second potential weakness of classification methods, related to the first, is that the actual group structure identified - whether 'genuine' (i.e. reflecting some external reality) or not - may be heavily dependent on the specific clustering algorithm used: "...different methods will often yield different groupings since each implicitly imposes a structure on the data." (Webb 2002, p. 361). The same objection is raised by Lavine, though his comments relate not only to cluster analysis but also cover two other group-seeking methodologies (factor analysis and multidimensional scaling): "One analysis may appear to show four well-distinguished groups but another may appear to show eight, three, or none at all." (Lavine 2008, p.84). From these two limitations - the unthinking mechanical location of possibly spurious structure, and the sensitivity of results to the precise algorithm used - a third weakness of classification follows. Since cluster membership is (for the reasons just given) potentially imprecise, the use of cluster schemes in inferential statistics is questionable. For example, if a formal test is conducted to determine whether the mean value of some quantity differs between cluster *A* and cluster *B*, the result of that test is arguably meaningless because the classification of objects into *A* and *B* is a mere mathematical abstraction reflecting the vagaries of the clustering algorithm used. A different clustering method may have assigned a member of *A* to *B*, and vice versa. Consequently, a formal test or statistical model which treats cluster membership as a measurement or precise attribute is unsafe. For this reason, clustering is often presented as an approach appropriate to exploratory data analysis (EDA) but not to formal inference. For example, Venables & Ripley (2002) present cluster analysis as a branch of exploratory multivariate analysis, and argue that visualisation methods are often far more effective for identifying groupings in data. The characterisation of clustering as a component of EDA is also argued by Webb: "Clustering is fundamentally a collection of methods of data exploration... ...The results of a cluster analysis may produce identifiable structure that can be used to generate hypotheses..." (Webb 2002, p. 361). Similarly, Jain *et al*. highlight the exploratory nature of clustering: "clustering methodology is particularly appropriate for the exploration of interrelationships among the data points to make an assessment (perhaps preliminary) of their structure." (Jain *et al*. 1999, p. 265). These objections to clustering as a basis for statistical inference are obviously potentially serious within the context of the present study. However, it is argued that they may be plausibly refuted on grounds which are now set out.

The use of clustering to generate inputs to formal statistical testing and modelling in this study may be partially defended on the grounds of *precedence*. Irrespective of the considerations outlined earlier in this section, cluster structures have been extensively used as the basis for inferential analysis, not least in the social sciences. Considered in isolation, this argument is not completely convincing (indeed, a "devil's advocate" response might present this simply as further proof of Cormack's assertion, cited earlier, that classification is often used 'irrelevantly and unjustifiably'). Nonetheless, by providing evidence of the prior use of classification for drawing formal inferences in social science research, it may be argued that the present study falls within an established (and largely unchallenged) tradition. For consistency, the examples now cited are the set of four diverse, recently-published papers which were presented earlier as exemplars of the use of classification methods in the social sciences.

Sabbe *et al*. (2008) used cluster analysis to group Flemish schoolchildren according to their patterns of activity and eating habits. Having done so, inter-cluster differences in gender, socioeconomic status and the prevalence of excessive body weight were investigated via chi-square tests. This study therefore provides an example of one of the classical techniques of statistical inference (the chi-square hypothesis test) being applied to a set of contingency tables in which one of the two classifying factors was derived not by measurement but purely via cluster analysis. Kuehne *et al*. (2008) used clustering to investigate factors influencing irrigation practices among Australian farmers and landowners. Between-cluster differences in a range of characteristics (e.g. length of time spent in the farming sector, land leasing / selling activity) were then assessed via formal methods (chi-square tests and Cramer's *V*). A study of drug use and sexual HIV risk conducted by Schonnesson *et al*. (2008) assigned a sample of HIV-seropositive African-American crack cocaine smokers to one of three clusters. Again, formal inferential techniques (one-way ANOVA and chi-square tests) were applied to investigate between-cluster differences in a wide range of variables expressing factors such as sociodemographics, drug use, sexual behaviour, HIV concealment and psychological functioning. Finally, Berry (2008) used cluster analysis to group residents of a disadvantaged rural Australian community according to their patterns of social participation. The resulting four clusters were again subjected to formal investigations of inter-group differences (using chi-square tests and factorial analysis of variance) in a range of variables such as age, sex, psychological distress, educational level and employment status.

Studies such as these illustrate the use of cluster analysis in the social sciences not merely in an exploratory role (e.g. to generate hypotheses), but to create classifications which are

treated as having substantive meaning. In other words, the groups identified in these studies are considered 'real' in the sense that (say) sex, blood type or ethnic group are real. Once this is accepted, it becomes valid to use the group structure in formal analysis just as would be the case for sex, blood type or ethnic group. However, as stated earlier, while there is a substantial precedent in the social sciences for the principle that cluster solutions may be used to draw inferences about wider populations, this does not in isolation confirm that such an approach is valid. To fully justify the use of clustering in the present study, some more detailed argument is required.

In this study, cluster methods are used to group subjects according to their time-related experiences of a single factor (SEP, or exposure to dampness or air pollution). This contrasts with many applications of classification, which seek to group objects based on values of multiple variables. It is arguable that the use of clustering in the special sense of classifying individuals by *time-variant values of a single factor* is not materially different, except in the level of detail involved, from established methods which are unquestioningly accepted as falling within the mainstream of research into health inequalities. To illustrate, consider a previously-cited study by Davey Smith and colleagues which sought to assess the effect of lifetime socioeconomic position on risk factors for cardiovascular disease, on morbidity, and on mortality from various causes among a sample of around 6,000 men in the West of Scotland (Davey Smith *et al*., 1997). In this study, subjects' social position was established at three stages of life (childhood, entry to the labour market and the time of screening for admission to the study). This information was used to construct an indicator of lifetime social position, by summing the number of time points (minimum zero; maximum three) at which the person was considered subject to social disadvantage (defined as holding manual occupational class). The associations of this indicator with the factors of interest were then assessed via formal statistical modelling. Now, it is perfectly reasonable to argue that the indicator used by Davey Smith *et al*. for this study (although treated as an accumulated quantity) is, in fact, a crude form of clustering on time-variant measurements of a single variable (in this case, social position). Individuals are 'clustered' according to whether they experienced no disadvantage, a single period of disadvantage, and so on. If this is accepted, then the use of a similar approach in the present study (albeit to a finer level of temporal detail) gains methodological respectability. This argument acquires even greater force when presented in relation to a study by Hallqvist and colleagues which consciously replicated the three-stage sampling approach of Davey Smith *et al*., but used the resulting data to define actual *trajectories* of social position (e.g. manual ➔ non-manual ➔ manual) rather than

accumulated measures (Hallqvist *et al*., 2004). These trajectories were then related by formal modelling (logistic regression) to the outcome of interest. It is difficult to challenge the view that this represents an application of clustering which is not dissimilar to that used in the present study. Thus, the methods used in the latter are extensions of those which are accepted as entirely appropriate by acknowledged authorities in health inequalities research. Indeed, it may be argued that the present study potentially improves on the method of Hallqvist *et al*. by defining trajectories based on a greater number of individual sampling points, leading to more precise characterisations of subjects' experiences. Informally stated, the clusters in this study are less vulnerable to misclassification error than those of the Davey Smith / Hallqvist studies, because there is less of the unknown in the classification process. In the particular use made of clustering in this study, it may plausibly be maintained that Cormack's maxim "we want to cluster only if clusters exist" is demonstrably satisfied. Clusters in this restricted sense clearly *do* exist: some people will experience (say) social disadvantage (or dampness, or air pollution) early in life, others later. For some, the experience will be sporadic; for others, more continuous. This is undeniable, and cluster analysis in this case represents no more than a convenient way of grouping together those who are exposed to a common influence at broadly similar times in life.

One further indirect defence of clustering in formal analysis is that one of the criticisms of its use in an inferential (as distinct from exploratory or descriptive) role arguably applies (in a more subtle form) to another major class of statistical techniques. As acknowledged above, "[clustering] algorithms will place cases into clusters, even when real clusters do not exist." (Fielding 2006, p.47). That is, the mathematical vagaries of the method used and the objects analysed can conspire to suggest relationships which are in fact spurious. However, a very similar criticism may be levelled at the automatic predictor selection methods (such as backward elimination and stepwise) which are very widely used in regression-type analyses. When such methods are employed, the predictor set in the model is selected purely on the basis of the mathematical properties of the data, without taking any account of whether the resultant predictor-outcome relationships are intuitively plausible. Such methods are routinely employed in research without incurring major criticisms, and it seems unreasonable not to extend such leniency to the employment of classification techniques. This, taken in conjunction with the other arguments advanced above, suggests that the use of clustering in the present study is defensible.

***

The method developed to condense the detailed sequences of social position into a higher-level typology is described in the following sections. The process consisted of two stages. The first step was to derive a measure of *distance*; that is, a quantity indicating how 'different' each sequence was from the remaining sequences in the set. Expressed more formally, "Distance coefficients are functions which take their maximum values (often 1) for two objects that are entirely different, and 0 for two objects that are identical over all descriptors." (Legendre & Legendre 1998, p. 274). The second step involved grouping the sequences on the basis of these pairwise distances such that similar sequences (that is, those for which mutual values of the distance measure were small) were placed in a common group. The first of these steps is described in Section 6.3.4, while the second stage is discussed in Sections 6.3.5 to 6.3.7.

6.3.4 Grouping the sequences (iii): deriving a measure of statistical distance

Deriving a measure of proximity (that is, of distance or similarity between objects) is an essential first stage in cluster analysis: "All clustering algorithms begin by measuring the similarity between the cases [here, sequences] to be clustered… …it is also possible to view similarity by its inverse, the distance between cases, with distance declining as similarity increases." (Fielding 2006, p. 48). Describing the derivation of a measure of distance begins with the statement of some basic principles. First, it is useful to stress at the outset that it is the sequence or trajectory (rather than the individual subject) which is now the unit of interest. In the processes to be described, it is immaterial whether a specific sequence is observed for a number of subjects in the dataset, or for only a single individual. The number of instances of each sequence is irrelevant to the objective, which is to define in quantitative terms how dissimilar each sequence in any pair of sequences is from the other. Second, it is helpful to illustrate the scale of the task; that is, the number of individual distance values which need to be calculated for the data used here. For any set of n objects, the total number of pairwise distances is given by $(n(n\text{-}1)) / 2$. Consequently, for the 216 sequences of social position identified in the dataset, the total number of individual distance measurements to be calculated would be $(216(215)) / 2 = 23{,}220$. Finally, it is important to appreciate that the calculation of a distance measure is influenced by the nature of the data. As described in Section 6.1, the sequences of social position created for this study consist of individual elements (each representing social location at a specific age point) which are nominal

quantities with four possible values (non-manual, manual, non-employed[24] and Armed Forces). While the first three of these arguably exhibit an ordered relation (from 'high' to 'low' social location), the fourth does not comfortably conform to such a scheme. As a result, decisions had to be taken as to whether (for example) the distance between the non-manual and manual states should be considered greater or less than the distance between the manual and Armed Forces states. This crucial issue is discussed later.

The literature on statistical proximity measures is substantial, and a large number of such measures have been developed. Descriptions of the major proximity coefficients are given by Sneath & Sokal (1973), Gower & Legendre (1986) and Legendre & Legendre (1998). One such coefficient, appropriate for quantifying the dissimilarity between string structures such as the social sequences of Figure 6.1.3, is the Levenshtein distance. This quantity, which is "the most common string dissimilarity measure used in the literature" (Coggins 1999, p. 312), is defined in the *Dictionary of Algorithms and Data Structures* maintained by the US National Institute of Standards and Technology (NIST) as "The smallest number of insertions, deletions and substitutions required to change one string or tree into another" (NIST, 2009a)[25]. Mathematically, the Levenshtein distance is a *metric* in that it satisfies the four axioms of a metric (Baake *et al.*, 2006; Kruskal, 1999); for three sequences *A*, *B* and *C*, the Levenshtein distance d_L has the following properties:-

1) $0 \le d_L(A,B) < \infty$ (the property of *positivity*)
2) $d_L(A,B) = 0$ if and only if $A = B$ (*non-degeneracy*)
3) $d_L(A,B) = d_L(B,A)$ (*symmetry*)
4) $d_L(A,B) \le d_L(A,C) + d_L(C,B)$ (*the triangle inequality*)

A number of modifications and extensions to the basic Levenshtein distance have been proposed. For example, Wagner & Fischer introduced a modified form based on the longest common subsequence between two strings (Wagner & Fischer, 1974), while Fu & Lu suggested the use of different weights depending on where in the string insertions are performed (Fu & Lu, 1977). Some other variations are discussed by Coggins (1999). However, in the absence of any compelling reason to use one of these variants (which are in general tailored to cater for the peculiarities of string comparison tasks in specific fields such

[24] The non-employed state is loosely considered to represent 'unemployed' (see Section 6.2.3); however, the former term is retained throughout this chapter.
[25] Levenshtein actually defined two concepts of distance, one conforming to the NIST definition given here, the other restricted to insertions and deletions (i.e. no substitutions are permitted) (Kruskal, 1999). The first sense is used throughout this thesis.

as character recognition), it was felt that the original Levenshtein distance was appropriate for use in this study.

Levenshtein distances may be derived via a process known as 'optimal matching' (MacIndoe & Abbott, 2004); this was the technique used in the study to derive a measure of distance between pairs of individual sequences of social location. The adoption of optimal matching was justified by its pedigree as a recognised tool for analysing sequence data in social science research, this record of successful application in a context directly related to that of the study making it a natural choice. While optimal matching analysis (OMA) has historically been used in a variety of fields including computer science, molecular biology and speech processing (Abbott & Forrest, 1986), it has more recently been employed in the social sciences to analyse sequence data. One of the first published reports of the application of OMA to sociological research described an analysis of sequences of figures performed by English Morris dancers (Abbott & Forrest, 1986). More recent applications of the method in social science research have investigated careers of musicians active in 18th century Germany (Abbott & Hrycak, 1990), the order of adoption of social welfare programmes in different countries (Abbott & DeViney, 1992), career systems in large financial institutions (Stovel *et al.*, 1996; Blair-Loy, 1999), analysis of work-life histories (Halpin & Chan, 1998), patterns of lynchings in the Deep South of the United States (Stovel, 2001) and predicting successful transitions from school to work (McVicar & Anyadike-Danes, 2002). The paper by Halpin & Chan is noteworthy in that their study was conceived largely as a critical evaluation of OMA: "We apply optimal matching techniques to class careers… …as a means of exploring the utility of this sequence-oriented approach." (Halpin & Chan 1998, p. 111).

Explanations of the optimal matching algorithm in a social science context are available, both in outline (Halpin & Chan, 1998; Stovel, 2001) and in detail (MacIndoe & Abbott, 2004). A comprehensive software implementation of OMA is available in the package TDA (Rohwer & Potter, 2007) which was used in the present study[26]. Halpin & Chan usefully summarise the concept behind OMA thus:

> What OMA does is count how many substitutions, insertions or deletions ('elementary operations') are needed in order to turn sequence A into sequence B, or vice versa." (Halpin & Chan 1998, p. 112.)

[26] During the development of the study, an additional software solution for applying optimal matching became available in the form of the TraMineR package for the R software system (Gabadinho *et al.*, 2008). This offers substantially greater ease of use than TDA, but unfortunately became available too late for use in the study.

The basic idea may be illustrated using a simple example. Consider two sequences, each consisting of ten binary elements: S1 = `0100000000` and S2 = `0000010000`. These might represent whether a person was, or was not, exposed to some hazard at ten consecutive yearly points. One way to align these sequences (i.e. render them identical) is to substitute the `1` at the second position of S1 (reading from the left) with a `0`, then replace the `0` at the sixth position of S1 with a `1`. This process, which involves two substitution operations, may be shown diagrammatically as follows:-

```
        ~~1~~       ~~0~~
         ^          ^
S1:   0 0 0 0 0 1 0 0 0 0
      | ^ | | | ^ | | | |
S2:   0 0 0 0 0 1 0 0 0 0
```

The same result may be achieved by insertion and deletion operations as follows. First, delete the `1` at the second position of S1, thus reducing the length of S1 to nine elements i.e.:-

```
0 ~~1~~ 0 0 0 0 0 0 0 0      ➔

0 0 0 0 0 0 0 0 0 _
```

Then, insert a `1` between the fifth and sixth positions of the truncated string resulting from the previous operation i.e.:-

```
0 0 0 0 0 0 0 0 0 _      ➔

0 0 0 0 0 1 0 0 0 0      (= S2).
```

Using this approach, alignment of S1 with S2 is achieved by one deletion and one insertion.

What is not immediately obvious from the above account is how OMA preserves the crucial element of *temporality* i.e. recognition of (for example) the fact that, in the present application, possession of non-manual social status at (say) age 20 is conceptually different from holding the same status at age 50. Were ordering in time not of interest, the whole concept of sequences - that is, ordered chains of events or (as here) observed conditions - could be dispensed with, and it would be necessary only to count the number of instances of

non-manual, manual etc. status in each individual's life history. Since a key interest of this study was the ordered relations in time which are hypothesised as linking social position, exposure to environmental hazards and health, it was necessary to be fully confident that use of OMA would preserve the time ordering in the data. Reassurance on this crucial concern is provided by Halpin & Chan as follows:-

> "Note that each elementary operation [i.e. substitution, insertion or deletion; the latter two sometimes known collectively as *indels*] deals with a pair of units between two sequences, rather than the position of these units in relation to other units in their own sequences. In this sense, each elementary operation is blind to the temporal order of events. However, in comparing two sequences a string of these elementary operations is carried out, and it is this repeated execution of local elementary operations that carries the sequential and temporal information." (Halpin & Chan 1998, p. 112)

As stated by Halpin & Chan, OMA is based on the concept of counting the number of elementary operations required to achieve identity between a pair of sequences. A key aspect of the technique is that weights (known as 'costs') may be associated with resolving each of the possible types of mismatch which may be encountered when comparing any two sequences in the set. The concept of cost is an important feature of the optimal matching method, and is succinctly illustrated by MacIndoe & Abbott:-

> "...an analyst may have theoretical reasons to treat some elementary operations as 'more costly' than others. Different replacements can be weighted differently in accordance with this theoretically driven scheme. Consider, for example, a set of career sequences consisting of five hierarchically ordered jobs 1 to 5 within an organization. In this simple code, job 1 is an entry-level position, jobs 2 through 4 are intermediate positions, and job 5 is a senior vice president position. Instead of assigning one standard replacement cost, the analyst may elect to define a replacement cost matrix in which the replacement of sequence element 1 for element 2 is less costly than the replacement of element 1 by element 5... ...One important variation in applications of OM algorithms in social science, then, is the setting of these various costs..." (MacIndoe & Abbott 2004, p. 389)

The criticality of setting the costs associated with elementary operations in optimal matching is confirmed by other authorities (e.g. Halpin & Chan, 1998; Stovel *et al.*, 1996), one remarking that "The assignment of transformation costs haunts all optimal matching analyses." (Stovel *et al*. 1996, p. 394). In the present project, the setting of costs was particularly problematic because, as previously highlighted, the four possible states in the

state space (non-manual, manual, non-employed and Armed Forces[27]) are not linked by any firm ordered relation. While an ordering may broadly be assumed to apply to the first three of these, it is extremely difficult plausibly to quantify the intervals between the states (and hence the OM costs to be assigned). For example, is the substitution of manual for non-manual status to be deemed more costly (that is, indicative of greater distance) than the replacement of the manual condition by non-employment? Is the cost of resolving a non-manual to non-employed mismatch equal to the sum of the costs associated with resolving the two intermediate mismatches (i.e. non-manual to manual plus manual to non-employed)? Challenging as these decisions are, the relationships of the three quasi-ordered states with the Armed Forces condition are even less easy to quantify. Is replacing the Armed Forces state with the manual condition more or less costly than replacing it with non-employment? Given that OMA seeks to identify the 'cheapest' (least costly) set of elementary operations required to achieve identity between any two sequences, it was clear that the specification of costs could materially influence the calculation of inter-sequence distance values, and hence the typology ultimately arrived at.

Recognising the impossibility of creating a definitive justification for a single set of cost rules in the present application, it was decided to specify two alternative cost schemes, applying each to the OM process in turn and comparing the results. A precedent for this approach exists; McVicar & Anyadike-Danes (2002) compared the respective effects of two different cost schemes when using optimal matching to group young people on the basis of sequential experiences covering the school-to-work transition. Consideration was given to creating a greater number of cost schemes for the present study, but this was rejected due to the danger of generating a proliferation of outputs which might be both unmanageable and difficult to interpret. Because it was planned from the outset to evaluate more than one clustering method in creating a typology of sequences, the joint effect of multiple OM cost schemes in conjunction with multiple clustering methods could easily lead to a potentially bewildering array of competing typologies (i.e. n cost schemes X m clustering methods = nm typologies). To avoid such a proliferation, the number of cost specification schemes was restricted to two. In the first of these, all elementary operations were assigned the default values specified in the TDA software i.e. 1 for an indel, 2 for a substitution. Under this system, resolving a mismatch in a single element incurs a cost of two irrespective of whether it is achieved via insertion / deletion or via substitution. This cost regime is referred to in the

[27] The conflation of the manual, Armed Forces and non-employed states, which was applied when deriving an accumulated measure of social disadvantage (Section 6.2), was not enforced when constructing the time-dependent measure of social location.

following material as the *default* cost scheme. In fact, the relationship between the respective costs of indel and substitution operations is a crucial factor to be considered when devising alternative cost specification schemes: "If substitution cost is higher than the cost of two indel operations, the [OM] algorithm will never choose substitutions." (Rohwer & Potter 2005, p. 1 in Section 6.7.2.5).

In creating a second cost scheme, an attempt was made to relate costs to the social 'distance' which separated each pair of states. For example, the distance between non-manual and non-employed states was regarded as greater than that between the manual and non-employed states. In an attempt to approximately reflect these presumed differences (which do not lend themselves to precise quantification), the substitution costs applied in the second costing scheme used were as shown (in matrix form) in Table 6.3.1. This shows that, for example, the cost of substituting the non-manual state for non-employed status was set at 3, while the cost of replacing non-manual with manual was 1. Any substitution involving the Armed Forces state (which lies outwith the presumed ordering inherent in the other three states) was given a uniform cost of 2 (midway between the lowest and highest costs defined for other substitutions).

TABLE 6.3.1: Matrix of substitution costs applied in TDA software when deriving pairwise distances between sequences of socioeconomic position. Cell content is the cost of turning the row element into the column element.

	non-manual	**manual**	**non-employed**	**Armed Forces**
non-manual	0	1	3	2
manual	1	0	2	2
non-employed	3	2	0	2
Armed Forces	2	2	2	0

Having decided on substitution costs, the final step before proceeding to perform OMA was that of specifying the cost of indel operations. One option was to set the indel cost so high that the OM algorithm would never select indels. In other words, alignment of sequences would be performed solely by substitution. However this was felt unduly restrictive, as it would involve essentially disabling a major feature of the OM process. Consequently, it was decided that the indel cost should be set at half that of the highest substitution cost. Since the value of the latter was 3 (see Table 6.3.1), a global cost of 1.5 was set for indels. Under this approach, indels would never be selected as an alternative to a single substitution when resolving a mismatch for which the substitution cost was less than 3, but would be available as an option for the following individual substitutions:-

replacing non-manual with non-employed
replacing non-employed with non-manual

Because it was tailored to reflect conceptual distances presumed to exist in the data, this second system of costs is referred to in the material which follows as the *tailored* cost scheme.

***

Having defined two alternative cost schemes as described above, the original intention was to apply the optimal matching method (using each scheme in turn) to the sequences of social location, thus deriving distance values which would then be input to a clustering procedure (see Section 6.3.5). However, initial experimentation with this process highlighted features of the social trajectory data which led to that intention being modified. The issues involved are of some importance to the concept of classifying individual experiences which underlies much of this study, and are now considered in detail.

Discussion begins with reference to Appendix 1. This shows a cluster analysis solution for distance values which were obtained by applying optimal matching to the social position sequences using the default cost scheme[28]. The solution of Appendix 1 assigns each of the 216 unique trajectories of social position observed in the data to one of ten clusters, thus defining what is conceived as a typology of time-related social position, in which broadly similar patterns of experience are grouped into a common category. The numbers of individual subjects assigned to each cluster are shown in Appendix 1, and two features of the data are immediately apparent. First, the respective numbers of subjects assigned to each cluster are highly unbalanced, ranging from 109 individuals placed in Cluster 2 to two respondents in each of Clusters 7, 8, 9 and 10. This disparity reflects the obvious reality that some patterns of experience are by their nature likely to be more (or less) common than others. For example, the experience of downward mobility in this sample (broadly represented by Cluster 3) is rare; a similar finding was observed in studies by Hallqvist *et al.* (2004) and Mishra *et al*. (2009). A second feature of the data is that the sequence-to-cluster assignments generated by the clustering process are, in some cases, potentially questionable. For instance, sequences 148, 153 and 158 are allocated to Cluster 1 (which loosely

[28] The basis on which this solution was chosen is not described here, but in summary selection relied on values of two measures (the pseudo F statistic and the pseudo t^2 statistic) which are available in the SAS CLUSTER procedure (see Section 6.3.6).

represents respondents who enjoyed persistent non-manual status), but might arguably be better placed in Cluster 4 (which holds cases who experienced upward social mobility).

The first of these characteristics (markedly unbalanced cluster sizes) has specific implications for analysis: the sparseness resulting from the presence of groups with very small numbers can preclude the calculation of maximum likelihood (ML) estimates, thus constraining the use of techniques (such as logistic regression) which depend on ML estimation. This limitation, which is largely unavoidable given the existence of uncommon patterns of experience, is not considered further here but is discussed in Section 6.3.1 below. The second feature of Appendix 1 (uncertainty or 'fuzziness' in the sequence-to-cluster assignments) is clearly a problem, in that ambiguities in group membership may potentially affect the quality of inferences based on that membership structure. If a sequence is assigned to group *A*, but would arguably more logically belong to group *B*, the validity of comparing these groups in terms of (say) a health measure is suspect. One obvious response to this challenge is to manually adjust the groupings, re-assigning sequences to clusters where it appears defensible to do so. The problems with this approach are twofold. First, it is a purely subjective process: ten analysts would in all probability perform the re-assignment in ten different ways. Second, although this method would permit instances of blatant misallocation to be corrected, the difficulty of treating genuine fuzziness (that is, sequences which have a plausible claim to membership of multiple clusters) would largely remain.

In view of these difficulties, consideration turned to the possibility of applying some form of pre-processing to the original sequences of year-on-year social position, with the goal of simplification: that is, reducing the level of between-sequence variation (and so eliminating some of the observed fuzziness), while preserving the essential character of each sequence. An obvious objection to such an approach is that it sacrifices information; however, this may be countered by arguing that the concept of deliberately discarding information in the interests of improved analytical tractability is routinely accepted in many areas of statistics. For example, in principal components analysis, high-dimensional data consisting of (say) n variables are reduced to m principal components ($m < n$), under the expectation that the m linear combinations of the variables retain much (but not all) of the original information. Similarly, in time series analysis smoothing of the original values (e.g. by taking moving averages or running medians) is often applied to reduce irregularities and thus clarify the underlying behaviour of the data series. More generally, the reduction of continuous data values to dichotomous or ordinal variables for analysis is commonplace. In some such cases,

the reduction is grounded on established theoretical considerations (for example, defining hypertension based on whether certain specific blood pressure levels are exceeded). In other instances, reduction of the original continuous quantity is largely arbitrary. What these approaches all have in common is that they involve a trade-off between intentional loss of information and analytical or interpretative convenience. On this basis, the simplification in this study of the social position sequences is justified on the grounds of extensive precedence.

While there are a number of ways in which such simplification could be achieved, the solution adopted was to extract and retain the values at every fifth year in the social location sequences from the age of 15 onwards (i.e. ages 15, 20, 25… …60; ten datum points in total). This may be viewed as a simple form of data reduction: the 46 individual 'observations' making up each sequence are condensed into ten, under the loose assumption that the discarded values centred around each retained point would, in the main, be the same as that of the retained point. Visual inspection of Appendix 1 broadly supports this assumption: frequent transitions in social location are relatively rare, so the value at retained year *Y* (say, age 25) will in most cases represent with reasonable accuracy the values of the discarded years (ages *Y*-2, *Y*-1, *Y*+1 and *Y*+2). Thus, the desired goal of simplification is achieved while preserving the main characteristics of each sequence.

After retention of data for every fifth year, the original sequence set of 216 unique social sequences (which would have generated a total of 23,220 pairwise distance values), was reduced to 122 unique trajectories (7,381 distance values). These reduced data were then subjected to the optimal matching process, using the two different OM cost schemes defined earlier. This process yielded two sets of values representing the statistical distance or dissimilarity between each pair of five-yearly sequences of social location found in the data. These values are most easily visualised in matrix form (see Figure 6.3.1 *[next page]*), and in fact the distance data generated by TDA had to be reshaped into this form for input to the clustering software used (the CLUSTER procedure in SAS software Version 9.1). The further processing of these inter-sequence distances to create a typology of time-dependent social position is described in the next two sections.

FIGURE 6.3.1: Generic representation of pairwise distances between sequences of socioeconomic position (matrix is symmetrical about the leading diagonal, so only the lower left triangle is shown).

	sequence 1	**sequence 2**	**sequence 3**	**sequence 4**	sequence *N*
sequence 1	0				
sequence 2	distance 1	0			
sequence 3	distance 2	distance 3	0		
sequence 4	distance 4	distance 5	distance 6	0	
sequence *N*	...	...	...	...	0

6.3.5 Grouping the sequences (iv): selection of clustering methods

The previous section described the derivation of two sets of measures representing the distance (or dissimilarity) between every pair of sequences of five-yearly social position observed in the data. In the present section, the first set of such measures (those derived using the default TDA cost values) is referred to as 'DM1' (for D*istance* M*atrix* 1). The second set (obtained using tailored cost values) is identified as 'DM2'. As previously explained, the goal of the processes described in this section was to use these distances to group the 122 unique sequences of social location into a typology of time-dependent social position which was sufficiently concise as to be usable in analysis. The two sets of distance values (DM1 and DM2) were in a sense competing with each other to be selected for use in this final typology, the construction of which constituted a major element of the whole project.

The general method used to reduce the full set of observed sequences was that of cluster analysis (see Section 6.3.2). Detailed accounts of cluster analysis, presenting the associated mathematics, are provided in many standard statistical texts (e.g. Johnson & Wichern, 1992; Timm, 2002; Hardle & Simar, 2007) and papers (for example, a thorough account of the core concepts and techniques is given by Jain *et al.* (1999). There is little merit in transcribing or paraphrasing such material here. Moreover, even a condensed account of the main clustering techniques would run to many pages and include numerous mathematical formulae. Recognising this, cluster analysis is for the purposes of this discussion treated largely as a 'black box' which accepts input in the form of distance values and generates the desired output: a cluster solution intended for use in further analysis as a typology of time-related social position. However, in view of the large number of clustering algorithms available, it is necessary to justify the selection of those actually used for this phase of the study.

As stated in the previous section, it was decided that multiple clustering methods would be applied to the data, and the results compared. Individual clustering algorithms tend to

exhibit particular strengths and weaknesses (for example, the complete linkage method is strongly affected by the presence of outliers in the data [Milligan, 1980]). In the absence of a compelling reason to use any specific clustering method, it was felt that reliance on a single method would be imprudent, since it would be impossible to estimate the extent to which the results were influenced by the peculiarities of that method. At the same time, it was clearly not practical to perform a full comparative assessment of several methods: this would constitute a substantial research effort in its own right. By way of a compromise, it was decided that two clustering methods would be implemented, each being applied to both of the sets of distance measures (i.e. to DM1 and to DM2). Thus four competing typologies of social position would result, and from these competing typologies a final selection would be made. Selecting the two clustering methods to be applied was challenging, as is frequently the case in cluster analysis: "The most important problem facing an investigator with data he would like to examine by clustering methods is that of which method to use." (Rand 1971, pp. 847-848). The task of selecting an appropriate method is complicated by the range of methods available: as long ago as 1979, at least 100 different clustering methods had been developed or proposed (Edelbrock, 1979). While "there is no optimal [clustering] method" (Timm 2002, p. 533), the method used is one of the key factors which may influence the results (i.e. the cluster solution which emerges):-

> "The analysis depends on the amount of random noise in the data, the existence of outliers in the data, the variables selected for the analysis, the proximity measure used, the spatial properties of the data, and the clustering method employed." (Timm 2002, p. 533)

A considerable literature relating to the evaluation of clustering methods exists (e.g. Rand, 1971; Edelbrock, 1979; Milligan, 1980) but it was not considered feasible to review this in depth for the present project. Rather, after consulting a small number of key sources, selection of two clustering methods was made on the basis now outlined. First, since there was no indication of how many clusters might be expected to occur in the data, it was decided that hierarchical methods (Johnson, 1967) were appropriate. Hierarchical methods offer the major advantage over alternative (non-hierarchical) methods that it is not necessary to know or estimate the number of clusters in advance of the analysis:-

> "In most real life clustering situations, an applied researcher is faced with the dilemma of selecting the number of clusters or partitions in the final solution... ...Virtually all clustering procedures provide little if any information as to the number of clusters present in the data. Nonhierarchical procedures usually require the user to specify this parameter before any clustering is accomplished and hierarchical methods routinely

> produce a series of solutions ranging from *n* clusters to a solution with only one cluster present (assume *n* objects in the data set).". (Milligan & Cooper 1985, p. 159)

Hierarchical clustering techniques may be classified as either agglomerative methods or diversive methods (Timm, 2002). In the former, clustering begins with each object (here, each sequence) assigned to its own cluster, these then being progressively combined until all objects are grouped in a single cluster. In diversive methods, the reverse approach is applied: all objects are contained in a single cluster which is successively split until each object is assigned its own cluster. For both sets of methods, the 'correct' or optimum number of clusters (determined via statistics which are discussed later) is assumed to lie somewhere between the starting and final cluster configurations. While a large number of distinct hierarchical clustering methods have been developed, the method of average linkage was found in a comparative evaluation to be among the best in terms of its ability to recover a known cluster structure when the presence of error or noise in the data was simulated (Milligan, 1980). Average linkage (which is an agglomerative method) was therefore selected as one of the two clustering methods to be used. The second clustering algorithm chosen for use was the centroid method (also an agglomerative algorithm). Selection of the centroid method was justified on the grounds that it has been demonstrated to be more robust to outliers than most other hierarchical methods (Milligan, 1980). As described above, each of these two clustering methods was applied to both of the distance matrices DM1 and DM2, resulting in four candidate cluster solutions of which one was selected for further analysis. The methods were applied via the CLUSTER procedure in SAS software, Version 9.1. The process for arriving at a preferred cluster solution for adoption in further analysis is now discussed

6.3.6 Grouping the sequences (v): identification of an optimum clustering solution

As explained above four clustering schemes were candidates for adoption, these schemes being generated respectively from

i. distance data matrix DM1 processed via the average linkage method; this analytical scenario is referred to hereafter as 'TC1' (from T*ypology* C*andidate* 1)
ii. DM1 processed via the centroid method ('TC2')
iii. DM2 processed via the average linkage method ('TC3')
iv. DM2 processed via the centroid method ('TC4')

Selection of the preferred solution was a two stage process. First, the optimum clustering scheme within each distance matrix / method combination had to be identified. This essentially involved locating the stage in the progressive agglomeration of clusters which would yield a cluster structure characterised by the minimum overlap (or greatest separation) between clusters, while also explaining an acceptable proportion of the observed variation in the distance values. Then, these four optimal solutions had to be compared to decide which was overall 'best'. These stages are now described in turn.

Numerous indices or procedures aimed at identifying the optimum solution in hierarchical cluster analysis have been developed or proposed, and their characteristics investigated. For example, a simulation study by Milligan and Cooper compared the performance of 30 such procedures (sometimes termed 'stopping rules') in recovering a known cluster structure (Milligan & Cooper, 1985). Two such indices which are available in the SAS CLUSTER procedure (the pseudo F statistic and the pseudo t^2 statistic) were used to identify the optimum cluster solution for each of the four analytical scenarios examined. These indices were used in conjunction with the R-square values calculated for each step in the cluster history, to ensure that the solution adopted would explain a meaningful proportion of the total variance in the distance data. A 'best' solution was selected on this basis for each distance matrix / clustering method combination, and the characteristics of these solutions are summarised in Table 6.3.2.

TABLE 6.3.2: Characteristics of clustering solutions derived from data representing statistical distance between sequences of socioeconomic position.

distance matrix / clustering method	**number of clusters**	**% of variance explained (R^2)**
TC1 (DM1 / average linkage)	13	84.1
TC2 (DM1 / centroid)	13	82.7
TC3 (DM2 / average linkage)	16	79.8
TC4 (DM2 / centroid)	17	74.0

Next, a decision had to be reached as to which of these four cluster schemes should be adopted for use in further analysis. The first stage in selecting a preferred scheme was to perform a comparison of the degree of similarity between the cluster solutions generated by TC1 and TC2, thus gaining some insight into the apparent influence of the clustering method used. This comparison identified little evidence of gross differences between the respective cluster schemes: the assignment of sequences (and hence of subjects) to clusters was broadly similar for both schemes. Overall, it appeared that the choice of clustering algorithm had, for this set of distance data, exerted relatively little influence on the cluster structure

identified. However, a comparison of solutions TC3 and TC4 indicated marked differences. For example, the three largest clusters in TC3 contained a total of 236 subjects (representing 80.3% of the total sample), while the three largest clusters of TC4 held no fewer than 273 subjects (= 92.9%). It therefore appeared that, for this set of distance data, the application of two different clustering algorithms had generated dissimilar cluster solutions. In deciding which solution should be considered preferable, it was clear that on purely pragmatic grounds the solution generated by the centroid method (analysis TC4) would in fact be largely unusable for the purposes of the study due to the extremely unbalanced distribution of subjects across clusters exhibited by this solution. As stated, this scheme assigned 93% of all subjects to one of just three clusters, thus effectively concealing many differences between individual patterns of social experience. On this basis, TC4 was eliminated from further consideration.

At this point, the three cluster schemes generated respectively by TC1, TC2 and TC3 remained candidates for adoption. Since the two former solutions had previously been judged to be very similar, the choice initially reduced to that of TC1 *or* TC2 *vs.* TC3. In deciding between these two alternatives, selection was guided by two considerations, namely (i) the anticipated usability of the cluster structure (in essence, the extent to which the structure contained large numbers of small clusters which would give rise to difficulties in subsequent analysis), and (ii) the degree to which the respective clustering schemes were interpretable (that is, whether members of Cluster *N* appeared to display identifiable common characteristics). Of these, the first was found to provide little guidance for further narrowing the choice. The respective numbers of subjects assigned to each cluster for the three remaining candidate solutions are shown in Table 6.3.3 *(next page)*, from which it is apparent that the general pattern of assignment is similar across all three solutions, being characterised by:

- two large clusters, together holding *c*. 70% of all subjects
- two further moderately sized clusters (Clusters 3 and 6 in TC1 / TC2, Clusters 3 and 5 in TC3) together accounting for around a further 20% of subjects
- a number of small clusters, each holding between 4.4% and 0.3% of subjects (representing respectively 13 individuals and one individual)

Clearly, none of the three solutions is ideal from a purely analytical point of view in that all are characterised by unbalanced cluster sizes. However, it was accepted that these clustering

schemes reflect the fundamental properties of the data (that is, that certain patterns of social experience are more common than others).

TABLE 6.3.3: Proportions of total number of subjects assigned to Cluster *N*, for cluster solutions generated by TC1, TC2 and TC3.

cluster	% of total subjects: TC1	% of total subjects: TC2	% of total subjects: TC3
1	32.0	36.4	36.7
2	37.1	37.1	34.4
3	13.3	12.6	9.2
4	1.7	1.7	4.4
5	0.7	0.7	7.1
6	7.1	7.1	0.7
7	3.7	0.7	1.7
8	1.4	1.4	1.0
9	1.4	0.7	1.0
10	0.7	0.7	1.4
11	0.3	0.3	0.7
12	0.3	0.3	0.3
13	0.3	0.3	0.3
14	N / A	N / A	0.3
15	N / A	N / A	0.3
16	N / A	N / A	0.3

Since the criterion of analytical usability provided no obvious basis for differentiating among the three remaining candidate solutions, the final choice of a scheme for representing social position in further analysis was made on the grounds of the interpretability and intuitive appeal of each scheme. This was, of course, not ideal: the overall objective of the fairly complex process leading to the creation of these clustering schemes was to eliminate subjectivity, as far as was feasible, in constructing one of the key measures required for this study. That this proved not to be wholly possible was disappointing, but reflected the reality that the application of clustering methods had ultimately resulted in the creation of three competing classification schemes which were broadly comparable. As a result, the selection of a final measure required the application of subjective judgement.

Further examination of the three remaining solutions led to the conclusion that the cluster scheme created by TC3 could justifiably be eliminated from consideration, this decision being taken on the following grounds. First (though really a usability criterion rather than a point of interpretation), the TC3 scheme contained three more of the analytically inconvenient single-subject clusters than its two competitors (see Table 6.3.3). Second, the TC3 solution included a notable anomaly in that the respective memberships of two clusters appeared conceptually very similar, and it was not immediately apparent why the sequences involved should have been assigned to separate clusters. Third, a further anomaly evident in

one cluster of TC3 appeared particularly threatening from the point of view of the analyses planned for the study. This cluster included sequences which contained substantial contiguous blocks of non-manual status both at the beginning of the period examined and towards the end. Since one of the basic postulates in the study is that social position at different points in the lifecourse may exert differential effects on health, the assignment to a single group of sequences which included an appreciable representation of non-manual status (hypothesised as being 'good') at such widely different points clearly had the potential to obscure the very effects which were of most interest. While none of these considerations was individually decisive, it was felt that their joint influence was sufficient to exclude the cluster scheme generated by TC3 from further consideration.

At this point, there remained only the two clustering schemes derived from the distance data obtained using the default optimal matching costs implemented in the TDA software. As indicated, these two solutions exhibited considerable similarity. The degree of resemblance between the two schemes was such as to make a final choice between them challenging. After careful consideration, a final selection was made based purely on considerations of analytical usability. From Table 6.3.3, it is evident that the number of clusters with very small membership is slightly lower in TC1 than in TC2. Specifically, Cluster 7 contains 11 subjects in TC1 as against 2 in TC2, while Cluster 9 holds 4 subjects in TC1, compared to 2 in TC2. Consequently, TC1 was judged to be marginally superior to its competitor in respect of the number of analytically inconvenient small clusters, though the degree of difference is clearly trivial. On this basis, the cluster structure derived from analysis TC1 was selected as the typology of time-dependent social position which would be retained, albeit subject to some further manipulation (see next section).

6.3.7 Grouping the sequences (vi): manual modification of the clustering solution

The cluster scheme derived as described above was not considered entirely suitable for further use in its original form, on grounds which are now outlined. As is evident from Table 6.3.3, five of the thirteen groups in the scheme contained very small numbers of subjects (either one or two persons). While reflecting the reality of individuals' lives (that is, the fact that certain patterns of experience are by their nature likely to be uncommon), the presence of these small clusters presented a serious threat to the programme of analysis planned for the study. The intended investigations included the use of logistic regression and, as highlighted earlier, this class of analysis is vulnerable to sparseness in the data, leading to the non-availability of maximum likelihood estimates under certain conditions.

Indeed, it was this consideration which largely motivated the decision to group or condense the original detailed sequences of social location, rather than using them in analysis directly (see Section 6.3.1). Formally, the difficulty is the existence of "a nonunique maximum on the boundary of the parameter space, at infinity" [Albert & Anderson 1984, p. 1]), and the condition is especially likely to arise when sample sizes are small:-

> "The difficulties associated with complete and quasicomplete separation are small sample problems. With increasing sample size, the probability of observing a set of separated data points tends to zero, no matter what the sampling scheme." (Albert & Anderson 1984, p. 9)

It was recognised that the inclusion in the cluster scheme of the very small groups evident in Table 6.3.3 would give rise to complete or quasi-complete separation of data points. To circumvent these difficulties, it was decided that the five groups which contained either one or two individuals would be excluded from analysis. While the loss of these subsets of subjects (seven respondents in total) was regretted, it was viewed as essential if the planned analyses were to be successfully accomplished.

While the above amendment to the original cluster solution was forced largely by pragmatic considerations (i.e. to render analysis possible), one final manual modification of the grouping scheme was applied on conceptual grounds. As Table 6.3.3 shows, the scheme is dominated numerically by Cluster 2; this cluster consisted of individuals whose social status during the period examined was mainly manual. However, of the 109 cases assigned to this group, 43 recorded a 'pure' manual experience (that is, were deemed manual at all ten five-yearly points between the ages of 15 and 60). It was decided that these 43 subjects should be assigned to their own group, thus dividing the original Cluster 2 into (a) a set of 43 individuals who experienced persistent manual status, and (b) a second set of 66 subjects whose social location over time was mainly (but not entirely) manual. This subdivision was justified on the grounds that comparison of these two groups (for example, in terms of health) would permit any effects associated with the undiluted lifetime experience of manual social status to be identified. Consideration was given to applying the same process (i.e. isolation of the 'pure' experience) to those whose social location was persistently non-manual, but was rejected because this pattern of experience was relatively rare, being confined to 14 of the 94 subjects in Cluster 1.

After application of these two manipulations, the classification system representing subjects' patterns of time-dependent social position consisted of nine groups or clusters. Details of the final scheme are given in the *Results* (Section 11.3).

***

At this stage, the two measures of social position required for the study were available, namely (i) a measure of accumulated disadvantage (Section 6.2) and (ii) a condensed representation or typology of social position over time (described in the present section). The next chapter describes the creation of corresponding measures of exposure to residential hazards.

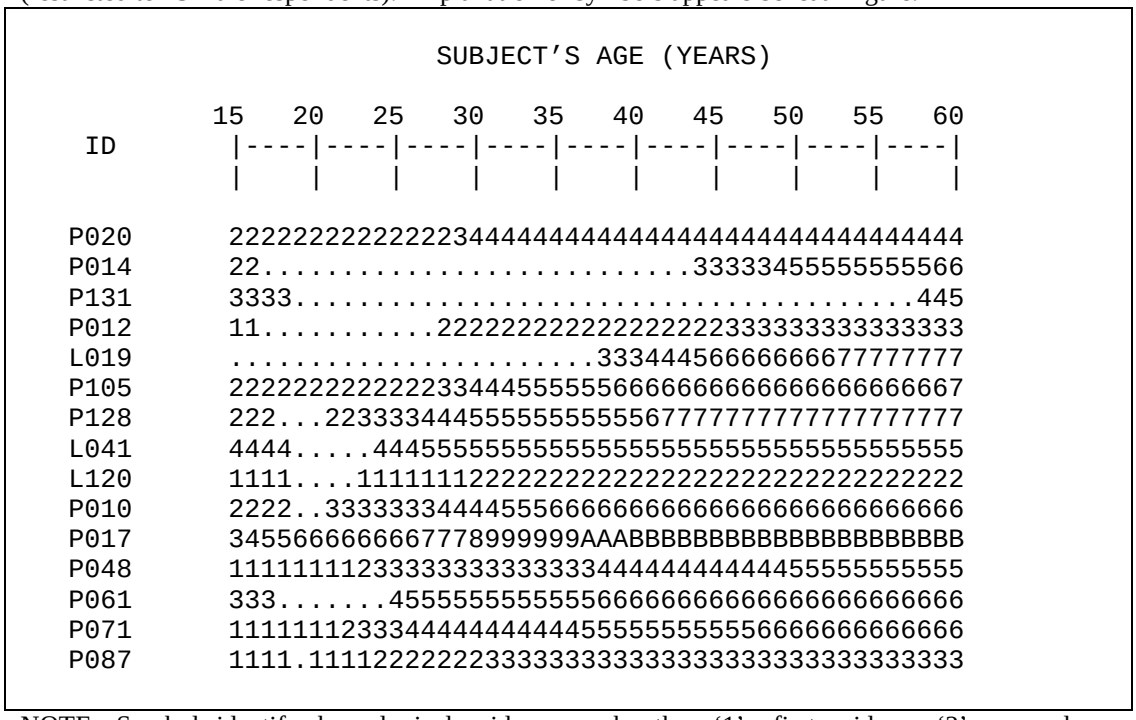
CHAPTER 7: METHODS (III) - MEASURES OF EXPOSURE TO RESIDENTIAL HAZARDS

7.1 Establishing the subject's housing history

Information on the respondent's exposure to residential hazards (specifically, dampness and air pollution) was derived directly from her / his housing history. The approach used in the dataset to represent the latter has been introduced in Section 5.3.2. In summary, the individual's housing history is represented by a family of variables which hold the start and end dates (expressed as calendar years e.g. '1932') of the subject's residence in each home lived in. Details are recorded for a maximum of 13 different dwellings. Multiple non-contiguous periods of residence at a single home are recorded separately in the data. For example, subject L027 lived in his first home from 1926 to 1939; again from 1941 to 1943; and finally between 1947 and 1948, moving to his second dwelling in the latter year. Because the subject's year of birth is provided, the start and end dates of periods of residence at an individual home may be used to identify the range of ages during which the person lived in that dwelling. Thus subject L027 (who was born in 1926) lived in his first residence from birth until the age of 13 years; from 15 to 17; and from 21 to 22. In order to provide the basic data structure needed to derive measures of residential conditions, the information on subjects' ages while living at each individual residence was initially used to construct sequences holding the number of the home (in chronological order i.e. 1, 2, 3 etc.) lived in at each year of life between the ages of 15 and 60 (see Figure 7.1.1 *[next page]*).

The sequences of residential history shown in Figure 7.1.1 incorporate a minor (and unavoidable) element of potential inaccuracy arising from the treatment of 'transition' years (that is, ages at which residence at home *N* ended, and residence at home *N*+1 began). This effect is best illustrated by an actual example. Subject P105 was born in 1924, and lived in his second home between the years 1937 and 1952 (i.e. the ages from 13 to 28). In 1952 he moved to his third residence. Because there is no indication of when in that year the house move took place, the year 1952 could plausibly be allocated to either of the subject's second or third residences. In fact, the computer routine used to translate residence periods (e.g. 1952 to 1954) into the range of ages during which the person lived in a specific house assigns the transition year to the later home. Thus in the case of P105, the year 1952 (i.e. age 28) is associated with residence 3 rather than residence 2. This element of ambiguity reflects an unavoidable limitation of the data, and is considered unlikely to introduce substantial levels of bias or inaccuracy in analysis.

FIGURE 7.1.1: Representation of subjects' residential histories across the age range 15 to 60 years (restricted to 15 male respondents). Explanation of symbols appears beneath figure.



NOTE: Symbols identify chronological residence number thus: '1' = first residence, '2' = second residence… …'A' = 10th residence, 'B' = 11th, 'C' = 12th, 'D' = 13th. Periods ('.') denote ages for which no residence information is available.

The sample of sequences shown in Figure 7.1.1 indicates that the housing history information exhibits substantial levels of missingness. The extent of the problem is illustrated by Table 7.1.1 *(next page)*, which shows the distribution of the number of individual years between the ages of 15 and 60 for which no identifiable residence is recorded. From Table 7.1.1 it is evident that a complete housing history is available for only 163 of 294 subjects (= 55.4%). Because the measures of exposure to residential environmental hazards available in the dataset are recorded as characteristics associated with each individual dwelling (see Section 5.3.2), the high levels of missingness in subjects' residential histories result in corresponding levels of missingness in the hazard data. That is, if it is not known which home the individual lived in at age *Y* years, it is impossible to determine whether s/he was exposed to poor housing conditions at that age. In response to this challenge, missing data in subjects' residential histories were under certain circumstances imputed according to a series of rules, considered to be conservative, which are now described.

TABLE 7.1.1: Distribution of number of individual years (between the ages of 15 and 60) for which no identifiable residence is recorded.

Number of missing years	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	163	55.44	163	55.44
1	43	14.63	206	70.07
2	21	7.14	227	77.21
3	13	4.42	240	81.63
4	10	3.40	250	85.03
5	5	1.70	255	86.73
6	5	1.70	260	88.44
7	3	1.02	263	89.46
8	5	1.70	268	91.16
9	6	2.04	274	93.20
10	2	0.68	276	93.88
11	3	1.02	279	94.90
15	1	0.34	280	95.24
16	2	0.68	282	95.92
17	1	0.34	283	96.26
18	1	0.34	284	96.60
19	1	0.34	285	96.94
20	1	0.34	286	97.28
22	2	0.68	288	97.96
23	1	0.34	289	98.30
24	1	0.34	290	98.64
27	1	0.34	291	98.98
30	1	0.34	292	99.32
39	1	0.34	293	99.66
46	1	0.34	294	100.00

Rule 1: single year missing

Where residence information was missing for only a single year, the missing value was replaced by the value for the previous year; that is, the person was assumed to have remained in her / his home for one further year. An example is provided by subject L136, whose residential history between 15 and 60 years contains a missing value at age 21 years *(see below - interpretation as for Figure 7.1.1)*:-

```
                    SUBJECT'S AGE (YEARS)

15   20   25   30   35   40   45   50   55   60
|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |
333333.444444444444444444444555555555555555
```

In this case the missing information for age 21 was completed under the assumption that the subject continued living in residence 3 for one further year, the revised (post-imputation) residence history for this individual becoming:-

```
                    SUBJECT'S AGE (YEARS)

15   20   25   30   35   40   45   50   55   60
|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |
3333333444444444444444444445555555555555555
```

The imputation can arguably be considered more reliable when the missing year is both preceded and followed by separate periods of residence in the same home; an example is provided by subject L080:-

```
                    SUBJECT'S AGE (YEARS)

15   20   25   30   35   40   45   50   55   60
|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |
2222222.22222223344445555555555555555555555
```

Note that this rule could not be applied under the following conditions:-

- the missing year is the first in the sequence (i.e. age 15); AND
- the value at the preceding year (i.e. age 14) is also missing

Application of this rule resulted in complete residence sequences being made available for a further 41 subjects (= 13.9% of the total number of cases).

Rule 2: two years missing (non-consecutive)

Where residence information was missing for two years, but the years involved were not consecutive, the missing value for each year was completed with the residence lived in at the previous year. This is essentially a multiple application of Rule 1. An example is provided by subject P122, whose original residence history exhibited missing values at age 19 and age 25:-

```
                    SUBJECT'S AGE (YEARS)

15   20   25   30   35   40   45   50   55   60
|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |
1111.11111.22223333333333344444444444444444
```

After imputation, the revised residence sequence for this individual became:-

```
                SUBJECT'S AGE (YEARS)

15   20   25   30   35   40   45   50   55   60
|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |
1111**1**11111**1**222233333333333344444444444444444
```

A further two subjects (= 0.7%) were recovered for analysis by applying this rule.

Rule 3: two years missing (consecutive)

Instances where residence information was missing for two years, but the years involved were consecutive (e.g. ages 19 and 20), were treated as follows. The first missing year was completed with the number of the residence preceding the pair of missing values, while the second missing year was assigned the number of the residence immediately after it. For example, the original residence sequence for subject P117 was:-

```
                SUBJECT'S AGE (YEARS)

15   20   25   30   35   40   45   50   55   60
|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |
2222..33334555555555555566666778888888888888
```

Following the application of Rule 3, this sequence became:-

```
                SUBJECT'S AGE (YEARS)

15   20   25   30   35   40   45   50   55   60
|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |
2222**23**33334555555555555566666778888888888888
```

Data for an additional 18 subjects (=6.1%) were recovered via application of this rule.

Rule 4: three years missing (non-consecutive)

In cases where information was missing for three years, but these three years did not form a single continuous block, each instance of missingness was handled as for Rule 1 or Rule 3. That is, where a single missing year was encountered the previous residence was held to apply; where two consecutive years were missing, the preceding and following residences

were applied respectively. A further two subjects (= 0.7%) were 'salvaged' by application of this rule.

Rule 5: three years missing (consecutive, preceded and followed by the same residence)

Where a contiguous block of three years was missing (e.g. ages 19 to 21) but the residence number was the same both before and after the three-year hiatus, the number of the 'surrounding' residence was presumed to apply. For example, the original residence history for subject L012 was:-

```
                SUBJECT'S AGE (YEARS)

15   20   25   30   35   40   45   50   55   60
|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |
1111...1111222222222222222233333333333333444444
```

but after application of Rule 5 this was modified to:-

```
                SUBJECT'S AGE (YEARS)

15   20   25   30   35   40   45   50   55   60
|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |
1111111111122222222222222223333333333333344444
```

Residential histories for a further four individuals (=1.4%) were deemed to be completed under this rule. Note that where a consecutive block of three years was missing, but the residences immediately preceding and following the hiatus were different, no imputation was applied as it was considered unsafe to do so.

***

After application of the above processes, the number of subjects for whom complete residential histories were regarded as being available (across the age range from 15 to 60 years) increased from 163 to 230 of 294 cases (= 78.2%). Consideration was given to extending the imputation process to cover more prolonged periods of missingness (i.e. situations in which four or more years' residence information was missing), but this was rejected. The reliability of the values assigned becomes more questionable as the imputation process grows increasingly aggressive, and it was considered that attempts to 'complete'

larger numbers of missing values would involve introducing unacceptable levels of uncertainty.

7.2 Constructing a detailed sequence of exposure to residential dampness over time

Having established the respondent's residential history as described above, the construction of sequences representing her / his exposure to residential dampness at each individual year was undertaken. To achieve this, the variables representing subjects' experience of probable and possible dampness at each residence (see Section 5.3.2) were used in conjunction with the residential histories. The process proved challenging for a number of reasons. First, the numbers of years of probable and / or possible exposure to dampness in residence *N* were in some cases lower than the total number of years during which the subject lived in that particular home. However the dampness variables hold only raw totals of the numbers of years involved, with no indication of when (chronologically) a period of dampness was experienced. Consequently, where the number of years' exposure to dampness was less than the total period of residence, it could not reliably be determined which years were characterised by damp. For example subject P135 lived in his second home between the ages of 24 and 32 (i.e. eight years), and reported the probable existence of dampness for four of those years. However, the dataset provides no means of identifying the specific age points at which this person was exposed to residential damp. To accommodate such situations, a general assumption was made throughout that any dampness was experienced in a single consecutive block at the beginning of the period of residence. On this basis, subject P135 was considered probably exposed to damp for four years from the age of 24 onwards. This approach was justified on the basis that the most likely explanation for a discrepancy between the total number of years spent in a dwelling, and the number of years' exposure to damp in that house, was eradication of dampness part-way through a period of residence[29]. While the reverse scenario - the onset of dampness in a previously dry dwelling – is not inconceivable (arising, for example, from a change in heating apparatus) it was viewed as less likely.

Application of the above approach was straightforward when the individual's residence at her / his *N*th home fell entirely within the age range of interest in the study (i.e. 15 to 60

[29] This is in fact the explanation advanced in the *Users Guide to the Dataset* for observed discrepancies between the total time spent in a dwelling and the length of exposure to dampness in that dwelling.

years). However, the assumption that 'partial' periods of dampness (that is, periods of shorter duration than the total length of residence) were located at the start of the subject's stay in a particular home had one subtle but potentially important consequence under the following combination of circumstances:-

i. the person's period of residence in a specific home fell partly before age 15; AND
ii. the number of reported years' exposure to damp for that home was lower than the total number of years spent in the dwelling

The issue may be illustrated by a hypothetical example. Suppose subject *X* spent the first 20 years of life in his first home, and that he was considered 'probably' exposed to residential damp for 10 of those years. Under the assumption that exposure to dampness was located at the start of the period in the home, *X* would be regarded as experiencing dampness for the first ten years of life, but would record no exposure to damp during the age range examined by the study.

Another challenge to creating histories of exposure to damp in the home arose when both probable and possible dampness were indicated for the same dwelling. This arose in only two instances:-

i. Subject L144 lived in his tenth residence for 27 years; he was possibly exposed to damp for 25 of these years, and probably exposed to damp for the remaining 2 years.
ii. Subject P128 stayed in his seventh home for 30 years, of which 24 involved probable exposure to damp while the remaining 6 were characterised by possible exposure to damp.

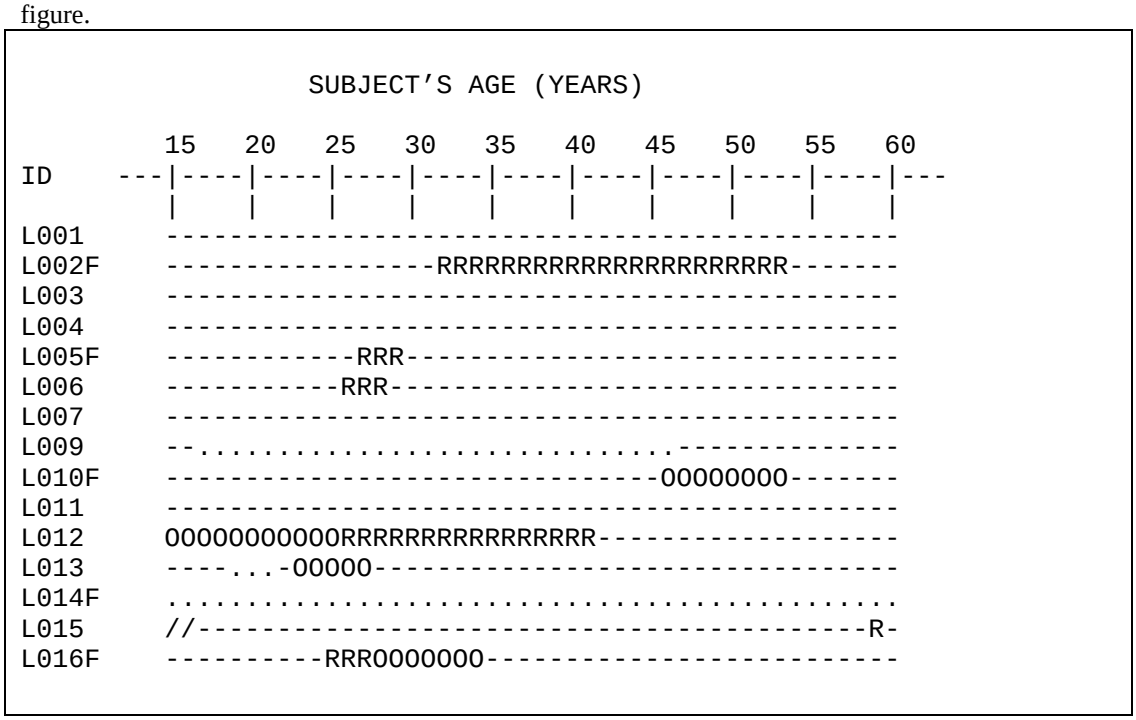
In these two instances, there was no method of determining which specific years of residence were associated with, respectively, the possible and probable exposures. Recognising this, it was decided that in these cases the more frequently-experienced state should be considered to apply for the subject's entire period of residence in the home involved. Thus, L144 was regarded as being possibly exposed to dampness throughout the 27 years spent at his tenth residence, while P128 was deemed probably exposed for the entire period of 30 years during which he lived at his seventh home.

Further complications arose as a direct result of the imputation processes described in Section 7.1. The effect of these was (for the $n = 67$ cases affected) notionally to 'increase' the number of years spent in certain residences. This, in turn, raised the issue of how the total number of years' exposure to dampness at the residences involved should be treated. Where the data indicated that the subject experienced damp throughout her / his stay in a specific dwelling, no difficulty was involved: the total number of years' dampness was simply incremented by the number of additional years of residence imputed for that home. However, challenges arose in the scenario where the individual was originally recorded as having lived in a home for n years, but was subject to dampness (either probable or possible) for m years ($m < n$). For example, where an additional two years of residence in one such dwelling were imputed (i.e. $n \rightarrow n+2$), should the number of years' exposure to dampness be incremented commensurately (i.e. $m \rightarrow m+2$), or remain unchanged? Clearly there was no wholly satisfactory solution to this dilemma. In the event the 67 cases potentially involved (i.e. those for which additional years of residence were imputed) were examined manually, and where the subject appeared originally to have been exposed to dampness for the greater part of her / his stay in the dwelling, the number of years' exposure to damp was increased by the number of additional years of imputed residence. Conversely, where the originally recorded exposure to damp was small relative to the total duration of residence in the dwelling, the number of 'damp' years was left unchanged.

After application of the imputations described in Section 7.1, and the specific adjustments described in the present section, the process to create a year-on-year record of exposure resulted in a data structure of the form illustrated in Figure 7.2.1 *(next page)*. In the figure, the individual's exposure to damp housing at each age point is defined as one of five states: probably exposed; possibly exposed; not exposed; unknown due to residence information being missing; and unknown due to dampness information being missing (i.e. the subject is known to live in Residence *N*, but dampness information is not recorded for that residence).

To complete creation of the sequences, the two states of exposure (probable and possible) were combined; that is, the subject was marked simply as being either exposed or unexposed at each age point. Complete sequences of exposure to dampness (i.e. data containing no instances of either type of unknown dampness status) across the age range 15 to 60 years were available for 220 subjects (=74.8%). The full sequence set is shown in Appendix 6.

FIGURE 7.2.1: Representation of subjects' exposure to residential dampness across the age range 15 to 60 years (restricted to first 15 respondents in dataset). Explanation of symbols appears beneath figure.



NOTE: Symbols identify dampness status thus: 'R' = damp probable, 'O' = damp possible, '-' = no damp, '/' = missing (residence known, but no dampness information), '.' = missing (residence not known).

7.3 Deriving a measure of accumulated exposure to residential dampness

Having established sequences of dampness experience as described above, the creation of a measure of accumulated exposure to damp was trivial, requiring only a summation of the number of years (minimum zero, maximum 46) during which the subject was exposed to either class of damp (i.e. probable or possible). The distribution of this quantity, and its summary properties, are shown in the *Results* (Section 11.4).

7.4 Creating a measure of time-dependent exposure to residential dampness

A measure of time-dependent exposure to dampness in the home was created via a method similar to that used earlier to derive a representation of time-related social position (see Section 6.3). The process was restricted to the n = 220 cases for whom complete dampness data were available. In outline, the steps involved were as follows. First, from subjects' full dampness histories, the dampness status (exposed or unexposed) was extracted at every fifth year (i.e. ages 15, 20, 25... ...60; ten datum points in total). Then, the method of optimal

matching was used to derive pairwise Levenshtein distances between the 43 unique patterns of five-year dampness experience which were present in the data. Because only two possible states were involved (i.e. dampness experienced *vs.* dampness not experienced), the process was markedly simpler than that developed for the social position trajectory data. In particular, the default cost scheme implemented in the TDA software package could be retained unchanged, thus avoiding the complexities resulting from the application of a tailored cost scheme. Finally, cluster analysis was used to group the dampness experience patterns (based on the pairwise statistical distances), yielding a typology of exposure to dampness over time.

Following the approach adopted for the social position data, two alternative clustering methods (average linkage and centroid) were applied to the dampness distance matrix. For each method, the optimum cluster solution was identified on the same basis as was done for the social location data (i.e. a local peak of the pseudo F statistic, co-inciding with a 'lagged' peak of the pseudo t^2 statistic). This approach suggested six-cluster solutions for both methods. The cluster solutions constructed via the two methods were in fact extremely similar, to the extent that no remotely persuasive case could be made for preferring one scheme over the other. Consequently, the solution derived via average linkage was (arbitrarily) selected for use.

As was the case with the grouped representation of social position (see Section 6.3.7), the dampness cluster solution derived as described above was not considered suitable for further use in its original form, and was subjected to manual amendment in two respects. First, one small cluster (holding only two individuals) was excluded from the scheme, because its presence would have generated problems when performing logistic regression analysis. Second, one very large cluster (containing 171 cases) was separated into (a) a subset of 129 respondents with no reported experience of dampness, and (b) the remaining 42 subjects (who were characterised by minimal exposure to dampness). Details of the final classification which resulted from application of these amendments are provided in the *Results* (Section 11.5).

7.5 Constructing a detailed sequence of exposure to air pollution over time

In addition to their use in constructing measures of dampness experience, the housing histories described in Section 7.1 also served as the basis for creating measures of exposure

to air pollution in the home. Following the approach adopted for the dampness data, the first step in this process involved the creation of detailed sequences representing subjects' year-on-year exposure to air pollution over the age range of interest. This task presented challenges, arising from the way in which exposure to pollution is defined in the dataset; the relevant extract from the *Users Guide to the Dataset* was reproduced earlier (in Section 5.3.2; item [b]). The scheme described there creates difficulties in constructing a year-on-year record of exposure to air pollution, in that an individual's exposure status may potentially change within her / his period of residence at a single home depending on whether the period involved crosses the definitional boundary represented by the calendar year 1960. The complexities inherent in such a definitional scheme may be illustrated by a simple example. Subject L025 lived in her sixth home between 1955 and 1985 – a total of 30 years which includes the 'critical' year 1960. For five of those years she is recorded as being possibly exposed to air pollution, and probably exposed for the remaining 25 years. However, these raw totals do not directly indicate which years (and, by extension, at which ages) the differing classes of exposure were experienced. To determine this key information, reference must be made to additional data elements which record the actual determinants of each dwelling's pollution status, namely:-

i. A variable indicating the type of area in which the *N*th residence was situated (one of: rural; small town; seaside town; suburban; urban; urban industrial)
ii. A binary indicator marking whether a factory was located within one mile of the *N*th residence
iii. A further binary indicator signalling the presence or absence of an A-road within 150 metres of the person's *N*th home

For L025's sixth dwelling, the values of these variables are respectively:-

- type of area: urban
- factory within one mile?: no
- A-road within 150m?: yes

From these values, it is possible to deduce the basis on which L025's experience of air pollution at this dwelling was assigned. Because the home was in an urban area (as distinct from an 'urban industrial' locale) with no local factory, years up to and including 1960 would attract a designation of possible air pollution. This accounts for the five years of

'possible' exposure to air pollution recorded in the dataset. The presence of an A-road (in conjunction with the urban location) within 150 metres means that any years after 1960 (here, the period until 1985) would be assigned a status of 'air pollution probable', accounting for the remaining 25 years spent at this dwelling.

In order to create a year-on-year representation of exposure to air pollution, the essential requirement was to translate (i) to (iii) listed above (which are attributes of the *dwelling*) into attributes associated with individual *age points* (e.g. age 25 years). However this process was not completely straightforward due to the need to recognise the 1960 boundary, additional complexity being introduced by (a) the provision for subjects to report multiple non-contiguous periods of residence at dwelling *N*, and (b) the imputation of additional notional years of residence under certain circumstances (see Section 7.1). The full process adopted to derive sequences of exposure to air pollution is now described:-

1) For each age point in the range from 15 to 60 years, the chronological number of the residence lived in at that age (1, 2, 3 etc.) was established.

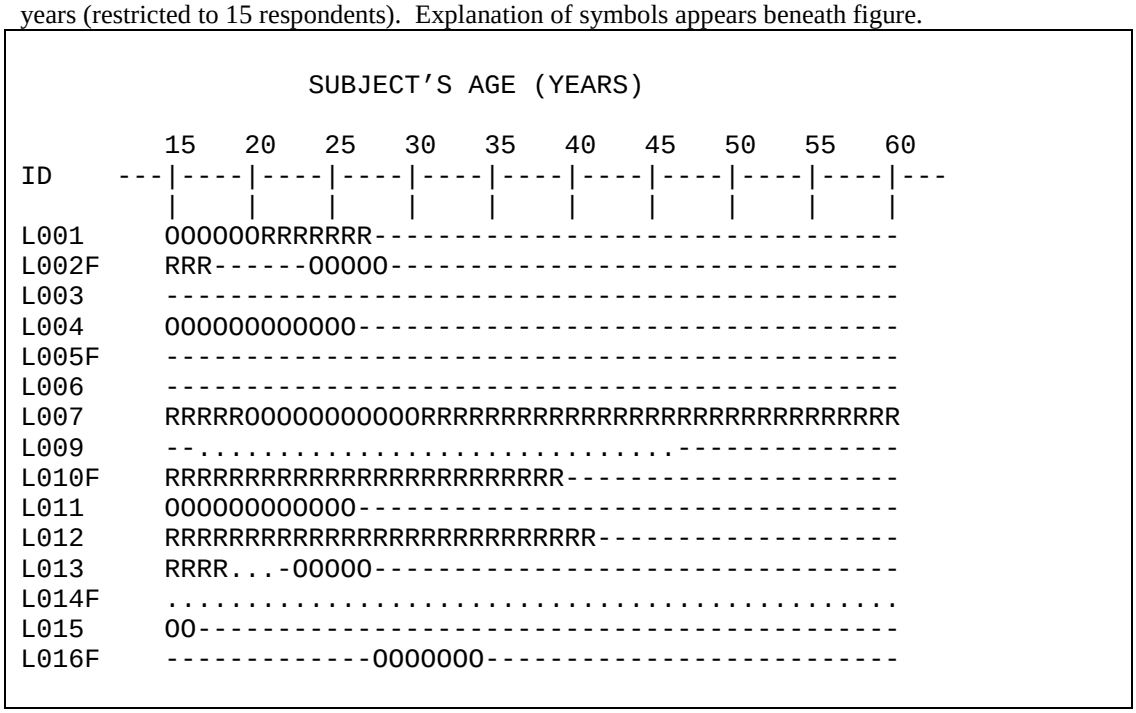
2) A test was applied to determine whether the calendar year associated with the age point was later than 1960.

3) If the calendar year involved was 1960 or earlier, the type of area associated with the residence was examined. If the area was classified as 'urban industrial', or if the area was 'urban' and was characterised by the presence of a factory within one mile, the subject was deemed to be exposed to probable air pollution at that age point. If the locale was urban, but no factory within one mile was recorded, exposure to air pollution at that age point was treated as possible. Under conditions other than those just outlined, no exposure to pollution at the age under examination was assumed.

4) Where the calendar year associated with age point *N* was 1961 or later, the type of area in which the dwelling was located was again determined. If the area was urban industrial or urban, and the existence of an A-road within 150 metres was indicated, the person was considered probably exposed to air pollution at that age. Where the area was urban or urban industrial but no A-road within 150 metres was indicated, possible exposure was inferred. Otherwise, an assumption of no exposure to air pollution at the age of interest was made.

The outcome of this process was a sequence of 46 datum points in the now-familiar form, defining the individual's exposure to air pollution at each year between the ages of 15 and 60 as one of: probably exposed; possibly exposed; not exposed; unknown due to residence information being missing; and unknown due to air pollution information being missing (i.e. residence known, but air pollution information not recorded for that residence). A sample of sequences is shown in Figure 7.5.1.

FIGURE 7.5.1: Representation of subjects' exposure to air pollution across the age range 15 to 60 years (restricted to 15 respondents). Explanation of symbols appears beneath figure.



NOTE: Symbols identify pollution status thus: 'R' = pollution probable, 'O' = pollution possible, '-' = no pollution, '/' = missing (residence known, but no pollution information), '.' = missing (residence not known).

Following the approach used earlier for dampness, final sequences were created by combining the probable and possible exposure states, the respondent simply being deemed either exposed or not exposed at each age point. A complete history of exposure to indoor air pollution (that is, data containing no instances of either type of 'not known' status) across the age range 15 to 60 years was made available for 215 subjects (=73.1%). Appendix 7 shows the full sequence set.

7.6 Deriving a measure of accumulated exposure to air pollution

Following the procedure adopted for the dampness information, a measure of accumulated exposure to air pollution was created by summing the number of years during which the respondent experienced either type of pollution (i.e. probable or possible). The properties of this measure are reported in the *Results* (Section 11.6).

7.7 Creating a measure of time-dependent exposure to air pollution

A representation of time-dependent exposure to residential air pollution was constructed via the same method as was applied to the dampness data i.e.

i. extraction of pollution status at every fifth year
ii. derivation of Levenshtein distances via optimal matching
iii. application of cluster analysis ('competitive' comparison of average linkage and centroid methods)

The process was restricted to the $n = 215$ cases for whom complete pollution sequences were available. Application of cluster analysis (evaluated, as before, by the pseudo F statistic and pseudo t^2 statistic) resulted in identification of an eight-cluster solution derived via average linkage and a more compact four-cluster solution obtained by the centroid method. No detailed comparison of these solutions is presented here, but two comments are appropriate. First, the centroid (four-cluster) solution was (unsurprisingly) characterised by a lack of within-cluster homogeneity, sequences which were markedly dissimilar being in some instances assigned to a common cluster. With only four 'slots' available to accommodate 39 unique patterns, this was obviously unavoidable. Second, the average linkage solution was notable for its relative ease of interpretation: of all the cluster schemes considered in the study, this is the one which was most intuitively 'sensible'. The average linkage solution was therefore adopted for further use as a typology of time-dependent exposure to air pollution.

As was the case for the grouped representations of social location and dampness exposure described earlier, the cluster solution depicting exposure to pollution was subject to manual adjustment. One small cluster (holding only two cases) was excluded, and one very large cluster (containing 144 respondents) was divided into (a) a group of 125 individuals with no

reported experience of pollution, and (b) the remaining 19 persons, who were characterised by minimal exposure to pollution. Details of the final scheme resulting from these amendments are presented in the *Results* (Section 11.7).

7.8 Deriving a measure of accumulated exposure to residential dampness or air pollution – 'total hazard load'

One final accumulated measure was constructed to represent the individual's total level of exposure (in years) to either dampness *or* air pollution. This quantity was derived by simply summing the respective number of years during which the subject experienced damp (see Section 7.3) or pollution (Section 7.6). Permissible values of the resulting measure ranged from a minimum of zero to a maximum of 92. This variable was regarded as representing the respondent's overall level of cumulative exposure to residential hazards – the 'total hazard load'. The properties of this measure are presented in Section 11.8 in the *Results*. Unlike the other accumulated variables discussed in this chapter, this quantity does not have a time-dependent (clustered) counterpart.

***

Thus far, the processes developed to create measures of social position, and of subjects' experience of residential hazards, have been described. The next chapter covers the treatment of exposure to occupational risks.

CHAPTER 8: METHODS (IV) - MEASURES OF EXPOSURE TO OCCUPATIONAL HAZARDS

8.1 Limitations of the representation of occupational hazards in the dataset

As outlined in Section 4.2, the study's interest in occupational exposures was restricted to determining whether such work-related hazards might operate as confounders in the association between social position and health. The variables in the dataset which record subjects' exposure to occupational hazards have been introduced in Section 5.3.4. While not immediately apparent from the brief account given there, the relationship in the dataset between the length of time for which an individual occupation was held, and the extent of the subject's presumed exposure to work-related hazards in that occupation, is complex and to some extent imprecisely defined. For example, in the case of exposure to occupational fumes and dusts the dataset records for each job (a) a score representing the number of years during which the individual was 'probably' exposed to the hazard in that job; (b) a similar score representing 'possible' exposure; and (c) details of the specific subtype(s) of hazard involved (e.g. coal dusts, solder / welding fumes). Up to three hazard subtypes are recorded for each job. Crucially, the score values (a) and (b) do not always simply sum to the numbers of years during which the respondent held the job. Rather, they may reflect adjustments applied when the data were collected to cater for non-continuous exposure (due for example to part-time working or seasonal patterns of exposure[30]).

Because of these adjustments, the variables representing work-related risks could not be used to create a detailed sequence of exposure similar to those constructed in the study for other hazards. Two examples will illustrate the difficulties involved. Subject L055 held her first job between the ages of 15 and 23 (i.e. for 8 years in total). A 'possible' fumes / dust exposure score of six years is recorded for this job, with an accompanying 'probable' score of zero years. There is thus a disparity between eight years spent in the job and six years of notional exposure. Similarly, subject L067 followed his first occupation between the ages of 14 and 19, and again between the ages of 21 and 23; he thus worked in this job for a total of seven years. However, the dataset indicates that he was considered exposed to the hazard (in this case physically arduous work) for five years. In such cases, it is clearly not feasible to define the person as being either simply exposed or free from exposure at specific age points during the period of employment. As an alternative, consideration was given to apportioning

[30] See the quotation from the *Users Guide to the Dataset* which is reproduced in Section 5.3.4 (item [a]).

the exposure score across the period during which the job was held, e.g. (in the case of L067) assigning five-sevenths of a notional 'year' of exposure to arduous work for each actual year of employment. However, this approach of assigning exposure *pro rata* would not yield a definitive exposed *vs.* unexposed contrast at each age point, which is needed to construct a sequence similar to those created in the study for exposure to social disadvantage and to residential hazards. In fact, defining exposure at individual ages as a quasi-continuous quantity in this way would result in complex structures unsuitable for processing with the analytical methods used in this project.

Recognising these difficulties, attention was turned to measures of accumulated exposure to work-related risks. In principle, representations of accumulated exposure could be constructed simply by summing the exposures recorded in each job to yield a grand total. Indeed, variables representing such totals - one for each of the three hazard types - are included in the original dataset (see Section 5.3.4). However, these are based on the individual's full working life rather than the common 'core' age range of 15 to 60 years which is used throughout this study. Attempts were made to create alternative measures of accumulation which were restricted to the core age range, but this task proved challenging. Difficulties arose in dealing with jobs which spanned either the lower or (more commonly) the upper boundary of the core age range (for example, a job which was first assumed at age 53 and held until the age of 65). In such cases, the total exposure score for the job (which may be less than the number of years for which the job was actually held) must be adjusted to reflect the fact that only part of the exposure took place within the period of interest. Implementing this approach proved difficult due to the wide variety of situations in which the phenomenon of boundary crossing could potentially be observed. For subject *A*, the fourth job (chronologically) might cross the 60^{th} / 61^{st} year boundary, while for respondent *B* the job in question might be his seventh. Moreover, a job which crossed the boundary was not necessarily the final one in the subject's employment history. For example, subject P077 took up his fourth job (which involved exposure to fumes) at the age of 53 and held it until he was 65; the job thus spanned the upper limit of the core age range. This individual then entered a fifth occupation (which did not involve exposure) between the ages of 65 and 67, followed by a sixth between the ages of 67 and 69, and a seventh from ages 70 to 74. The two later jobs also involved exposure to fumes.

Such complexities were compounded by the fact that each job could potentially involve multiple non-contiguous periods of employment. Thus, for subject *A* it might be the second

period of employment in his fourth job which crossed the 60 year threshold, while for subject *B* the boundary might be spanned by the third period of his seventh job. Considerable effort was devoted to developing a computer algorithm which would correctly adjust the recorded exposure values for these boundary-crossing effects. However, the large number of possible conditions and variations involved meant that the results were not wholly satisfactory. Eventually, it was decided that the effort required to develop and verify a fully robust method for calculating accumulated exposure within the 15 to 60 age range was disproportionate to the potential benefits. This decision took account of the fact that the exposure scores recorded in the dataset are themselves in some cases adjusted estimates (as discussed above). Consequently, even an algorithm which correctly catered for all possible combinations of conditions would not yield results which could be considered precise. Reflecting this decision, a simpler alternative approach to representing exposure to work-related hazards was adopted. This used the measures of total lifetime exposure, and is based on the argument which is now presented.

8.2 Representing approximate exposure to occupational hazards via binary indicators

Although the accumulated lifetime exposure totals in the dataset do not precisely represent subjects' exposure over the restricted 'core' age range of 15 to 60 years, it is arguable that they may be regarded as reasonable approximations to the latter. A person whose lifetime exposure to (say) fumes is substantial is likely to have also experienced high levels of exposure in the core age range. Similarly, individuals with little lifetime exposure to fume hazards will in general have experienced low levels of exposure over the core period. Moreover, in the special case of zero lifetime exposure, the corresponding level of exposure in the core age range ceases to be an approximation: it will (obviously) also be zero. The state of zero lifetime exposure to work-related hazards is in fact fairly common in the dataset, being observed for 46.9% of subjects for the fumes / dust hazard, 50.2% for arduous work and 52.9% for job demand / control stress.

Based on the above considerations, it was decided to use the original accumulated lifetime exposure totals in the dataset to create binary markers providing an approximate indication of whether the subject was, or was not, likely to have been exposed to appreciable levels of each hazard over the period from 15 to 60 years. Clearly, such a dichotomisation of the original hazard variables - which are quasi-continuous (integer) quantities - could be realised in many different ways. Consideration was initially given to setting the dichotomisation

boundary at the median value of the original variable (i.e. median or lower = not substantially exposed, above median = substantially exposed). However, for two of the three variables involved the median values are zero. For these variables, fixing the dichotomisation boundary at the median would effectively define a contrast between zero exposure and any exposure. This was felt to offer insufficient discrimination between trivial and high non-zero levels of exposure. As an alternative, it was decided that the 75th centile (third quartile) of each distribution would be used as the dichotomisation boundary, on the grounds that this would isolate a sizeable subset of cases whose level of exposure could be considered substantial relative to that of the remaining subjects. Under this scheme, respondents whose original (lifetime) recorded value for accumulated exposure to each hazard fell on or below the 75th centile were regarded as experiencing low exposure to that hazard within the core age range, while those whose observed values were above the 75th centile were treated as having received high exposure. The properties of the measures resulting from the dichotomisation process are summarised in the *Results* (Section 15.1). The use made of these three indicators (representing, respectively, exposure to fumes / dusts, arduous work and job demand / control stress) in the programme of analysis conducted for the study is discussed in the next section.

8.3 Occupational hazards as potential confounders

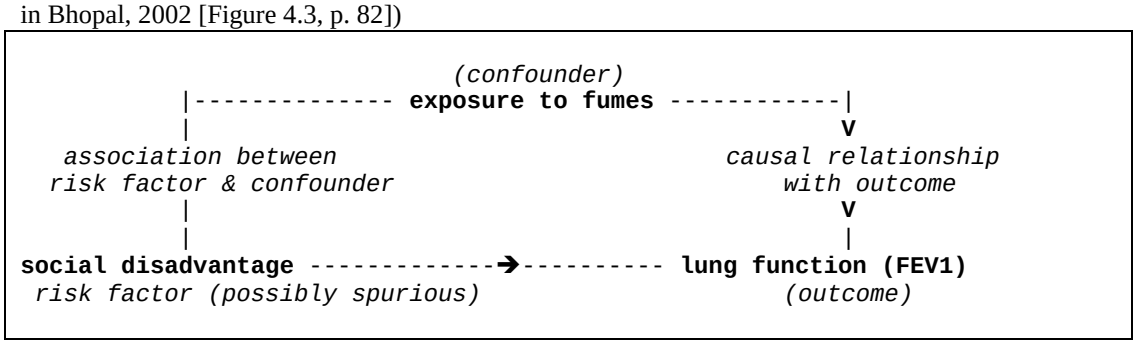
As explained earlier, the measures of occupational risk introduced in Section 8.2 were intended to permit a determination of whether work-related hazards might operate as confounders in the main relationships of interest. The concept of confounding is fundamental to epidemiology, and is routinely treated in textbooks devoted to that discipline (e.g. Bhopal, 2002; Koepsell& Weiss, 2003; Szklo & Nieto, 2007). A general definition of confounding is given by Bhopal thus:-

> "In essence, it [confounding] is the error in the estimate of the measure of association between a specific risk factor and disease outcome, which arises when there are differences in the comparison populations other than the risk factor under study." (Bhopal 2002, p. 79)

In the context of this study, it is reasonable to consider whether the main associations under investigation (e.g. that between social disadvantage [a risk factor] and standardised FEV1 [an outcome]) might be wholly or partly explained by a confounding effect of occupational risk (such as a hypothesised propensity for those in lower occupational positions to work in

potentially health-damaging jobs involving exposure to fumes). This scenario is shown schematically in Figure 8.3.1.

FIGURE 8.3.1: Hypothesised confounding effect of exposure to occupational fumes on the relationship between social disadvantage and lung function (based on the general representation given in Bhopal, 2002 [Figure 4.3, p. 82])



Testing and adjusting for the presence of confounding is a common challenge in epidemiological studies. One frequently recommended response is to perform multiple separate analyses, with the data stratified by the suspected confounder. Stratified analysis is recommended by Bhopal (2002), and its benefits are confirmed by Wallace: "Stratification… …adjusts for, or neutralizes, the effects of a confounding factor." (Wallace 2007, p. 22).

In the present study, stratification could at first sight be easily accomplished. The putative confounders (that is, the measures of occupational exposure described in Section 8.2) are binary quantities, and the application of stratification would thus merely involve performing each analysis twice (once for each level of the work-related hazard). However, this option was not feasible for many of the key analyses required. Some of the planned analyses involved the use of logistic regression, and difficulties arising from sparseness in the data (see Section 6.3.1) would be amplified considerably if the clustered groups were further divided by stratification. In fact, many of the planned analyses involving these time-dependent measures would not be feasible if stratification were applied.

However, even if stratification had been possible, there is no guarantee that its application would have fully elucidated whether the relationships of interest were distorted by the confounding influence of subjects' exposure to work-related risks. A more general problem in testing and controlling for confounding is that unless potential confounders are measured perfectly (i.e. with zero measurement error), and their association with the exposure perfectly characterised, bias resulting from confounding cannot be completely removed (Fewell *et al.*, 2007). Because such perfection of measurement is (obviously) unattainable in many cases, it

has been suggested that "effect sizes of the magnitude frequently reported in observational epidemiologic studies can be generated by residual and / or unmeasured confounding alone." (Fewell *et al.* 2007, p. 646). In other words, even where attempted correction for confounding is possible, the main effects observed after adjustment may be attributable to confounding influences which have been imperfectly measured.

Recognising the impossibility of comprehensively correcting for any potential confounding effect of work-related exposures, it was decided to limit the assessment of occupational confounding influences to three simple investigations. First, associations between the indicators of occupational exposure and the health outcomes of primary interest (i.e. those expressing aspects of cardio-respiratory health) were investigated. For each continuous health measure, its relationships with the three occupational indicators were assessed via Mann-Whitney tests. In the case of the binary health indicators, associations were determined via Fisher's exact tests. This process yielded, for each occupational hazard / health outcome combination, an estimate of whether the rightmost vertical association in Figure 8.3.1 was present, thus providing an outline indication of whether a confounding effect might potentially be in operation. Results from these tests are presented in Section 15.2 in the *Results*.

Second, the relationships among the three occupational exposure variables themselves were examined via Fisher's exact tests, with each pair of variables arranged in a contingency table e.g.:-

	arduous work: LOW	arduous work: HIGH
fumes / dusts: LOW	*n1*	*n2*
fumes / dusts: HIGH	*n3*	*n4*

The rationale behind this set of tests (three in total) was that they would permit a crude determination of whether the three specific occupational hazards recognised in this study appeared to be independent of each other, in the sense that the experience of hazard *A* did or did not co-exist with that of hazard *B*. This, in turn, would indicate whether 'work-related risk' should most appropriately be regarded as a single construct, or whether clearly separate effects were potentially involved. Results from these tests are given in Section 15.3.

Finally, the relationship of each occupational exposure variable to subjects' smoking history was assessed. The association between a dichotomous representation of smoking status (current / former *vs.* never) and each occupational hazard indicator was again assessed via a

Fisher's exact test (three tests in total). The objective of this group of tests was to estimate whether any effects of smoking on the main associations of interest did, or did not, appear to be inseparable from any potential influence attributable to the occupational exposures. For example, if it were demonstrated that individuals who were current or former smokers also tended to experience work-related fumes, this would indicate that the respective effects of these two factors (smoking and fume exposure) were effectively inseparable for analytical purposes. The results from this set of tests appear in Section 15.4.

It was recognised that the three sets of investigations described above did not provide a comprehensive assessment of whether exposure to work-related hazards might operate as a confounding influence on the main relationships of interest in the study. However, the rationale behind these investigations was that they would provide some degree of informal insight into whether such confounding effects were likely to be present, and would consequently be of value when interpreting the study's results.

***

Discussion now proceeds to describe the methods used to investigate the main associations of interest in the study.

CHAPTER 9: METHODS (V) – ESTIMATING ASSOCIATIONS

9.1 Summary of associations to be investigated

Assessing the three conceptual models outlined in Section 4.2 required that six *sets* of associations (each consisting of multiple individual associations) be investigated, namely:-

i. time-dependent SEP *with* health (required in: Model A)
ii. time-dependent SEP *with* time-dependent residential conditions (required in: Model A)
iii. time-dependent residential conditions *with* health (required in: Model A, Model C)
iv. accumulated social disadvantage *with* health (required in: Model B, Model C)
v. accumulated social disadvantage *with* accumulated exposure to adverse residential conditions (required in: Model B, Model C)
vi. accumulated exposure to adverse residential conditions *with* health (required in: Model B)

In the above list, the *time-dependent* measures are the grouped or clustered trajectories of SEP (see Section 6.3), exposure to dampness (Section 7.4) and exposure to air pollution (Section 7.7). These quantities, which are hypothesised to represent general patterns or typologies of experience over time, are referred to throughout this chapter as 'time-dependent measures' to distinguish them from individual (i.e. ungrouped) trajectories or sequences such as those illustrated in Figure 6.1.3. The *accumulated* measures in the list above are the quantities whose derivation is described in, respectively, Sections 6.2 (social disadvantage), 7.3 (dampness), 7.6 (air pollution) and 7.8 (total hazard load).

In the remainder of this chapter, description of the methods used to investigate the large number of individual associations involved is eased by grouping the health-related variables into a manageable number of conceptually related categories. In doing so, a clear distinction is made between those measures which express aspects of cardio-respiratory health[31] (and thus relate most directly to the theme of the study) and those which do not. In the following material, and in the *Results* chapters, health variables are grouped into four categories which are now defined.

[31] See the final paragraph of Section 5.3.3.

Physiological variables

This term is used to cover the three variables representing physiological parameters: systolic blood pressure (introduced as item [g] in Section 5.3.3); diastolic blood pressure (item [h]); and standardised FEV_1 (item [j]).

Clinical variables

This group of measures includes the four binary indicators marking the presence of, respectively; heart disease, lung disease, stroke and high blood pressure (see item [c] in Section 5.3.3).

Medication usage variables

This pair of variables consists of the dichotomous indicators of anti-hypertensive use (item [d]) and bronchodilator use (item [e]).

Secondary health outcomes

This final category includes all of the remaining health measures listed in Section 5.3.3; that is, those which do not specifically express aspects of cardio-respiratory health. There are ten such variables (all of them binary quantities), representing respectively the presence of long-standing illness (item [a] in Section 5.3.3) and of limiting long-term illness (item [b]); the experience of joint disease, diabetes, abdominal hernia, thyroid disease, duodenal ulcer, cancer and any other illness (item [c]); and the use of any prescribed medication (item [f]). This group of measures is not considered in detail, either in the present chapter or in the *Results*.

9.2 Analytical concerns: missingness and multiple testing

Before proceeding to outline the methods used to estimate the associations of interest, two important aspects of the investigation are introduced. First, it is emphasised that most of the variables featured in the analyses exhibit some degree of missingness (e.g. the representation of accumulated exposure to air pollution is available for only 215 of the 294 cases in the

dataset). Recognising this, a decision had to be taken as to whether analysis should be restricted to a fixed subset of subjects for whom complete data were available (i.e. no missing values existed for any of the measures involved in association sets [i] to [vi] above), or should exploit the maximum number of cases available in each specific analytical situation. For example, in assessing the relationship between accumulated disadvantage and systolic blood pressure, the analysis could be based either on those subjects who have nonmissing values for every measure, or on the larger number who have complete data for these two variables, but may have incomplete information for other measures. The use of a fixed 'core' of subjects is obviously attractive, because it ensures that all associations are estimated on precisely the same basis. However it was felt that limiting analysis to a static subset of individuals for whom complete information was available would exclude an unacceptably high proportion of what is, to begin with, an undesirably small dataset. Therefore, it was decided that each individual analysis should use the maximum number of cases available.

A second feature of the analyses is that the number of individual associations examined by the study was large. This leads to difficulties of interpretation, arising from the possibility of statistically significant results being obtained due to the random play of chance rather than to the existence of real effects. The problem of such Type I errors ('false positives') is widely recognised, and is commonly highlighted in texts devoted to statistical analysis (e.g. Armitage & Berry, 1987; Altman, 1991; Rice, 1995). Although most commonly discussed in the context of hypothesis testing, the challenge of interpretation also arises when the results of analysis are presented as confidence intervals around an estimated parameter (Katz, 2002). The applicability of the multiple testing dilemma to the latter is clarified by Stryer & Browner:-

> "…because a confidence interval is data-based, it cannot provide more information than a P value in the determination of the likelihood that a result represents a true positive. Instead, a 95 percent confidence interval represents the range of values that are consistent with the study result; any value outside the confidence interval would be rejected if P was less than 0.05." (Stryer & Browner 1994, p. 861)

The challenge has been extensively discussed by epidemiologists, and views both dismissive of the need to adjust for multiple test results (Rothman, 1990; Savitz & Olshan, 1995) and supportive of such corrections (Mills, 1993; Thompson, 1998) have been expressed. While no consensus on the correct response to multiple testing has emerged, the treatment of the topic by Rothman has proved influential, and forms the basis of the stance taken in the

present study. In essence, Rothman argues that no adjustment for multiple testing is needed on the following grounds. Adjustment for multiple tests is generally urged on the basis that 'chance' is the most likely explanation for observed effects; that is, a 'universal null hypothesis' underlies most events and associations which are identified in research. However, the default attribution of associations to chance is in fact open to challenge:-

> "Scientists presume instead that the universe is governed by natural laws, and that underlying the variability that we observe is a network of factors related to one another through causal connections. To entertain the universal null hypothesis is, in effect, to suspend belief in the real world and thereby to question the premises of empiricism. For the large bodies of data for which adjustments for multiple comparisons are most enthusiastically recommended, the tenability of a universal null hypothesis is most farfetched. In a body of data replete with associations, it may be that some are explained by what we call 'chance', but there is no empirical justification for a hypothesis that all the associations are unpredictable manifestations of random processes." (Rothman 1990, p. 45)

A less rarefied (but equally convincing) argument against adjustment for multiple testing is advanced by Katz, who applies a kind of *reductio ad absurdum* reasoning:-

> "…an individual comparison cannot 'know' how many other comparisons you have made. Therefore an individual association cannot be more or less likely due to chance based on how many other associations you have assessed… …If you favour adjusting for multiple comparisons, should you adjust for the number of comparisons you assessed in a single paper, or the number of comparisons assessed in a series of papers analyzing the same data set, or the number of comparisons performed during your career?" (Katz 2002, pp. 139-140)

A refutation of the need for multiple testing adjustments is made on similar grounds by Perneger, who asks (rhetorically) whether one should adjust for tests which were performed but not published, or should attempt to account for tests which may be performed in the future on the same data (Perneger, 1998).

It is of course possible to conceive situations in which an observed effect could reasonably be ascribed to chance; Rothman gives the example of an association between use of chewing gum and the occurrence of brain cancer. However, the main relationships postulated and tested in this study do not fall into this category. Rather, the associations of interest generally possess a degree of inherent plausibility, which weakens the interpretation that any significant result should automatically be considered suspect. Because of this, and on the basis of the arguments presented above, no attempt was made in the present study to adjust for the large number of tests and confidence interval estimations performed.

***

The methods used to assess the sets of associations listed in 9.1 above are now outlined. In the interests of clarity, the following material groups individual associations by *analytical regime*: all associations which were investigated by a common approach (which might involve either a single statistical technique, or a combination e.g. rank correlation accompanied by multiple regression) are considered together. This treatment avoids the necessity for repeated descriptions of the methods involved[32]. A total of six distinct analytical regimes were employed; these are now described in turn.

9.3 Analytical regime 1: initial investigation via rank correlation, followed by multiple regression

The first analytical regime was applied to investigate the following associations:-

i. accumulated disadvantage *with* the physiological variables
ii. accumulated exposure to residential hazards *with* the physiological variables

These sets of associations were initially assessed via rank correlation, the goal at this stage being to determine whether any monotonic relationship between each accumulated measure and each physiological parameter was present. That is, did greater (or smaller) levels of, say, cumulative disadvantage tend to co-exist with higher (or lower) values of the physiological measure? These relationships were then investigated in greater detail via multiple regression. Models were fitted in which each of the physiological variables was predicted by each cumulative measure, together with additional predictors representing sex and smoking status (the latter expressed as a binary contrast between current or former smokers and those who have never smoked). These models may be expressed algebraically in the usual form for a linear model; for example, the model fitted to assess the relationship between cumulative exposure to dampness and systolic blood pressure was:-

[32] In the actual presentation of findings in the *Results* chapters a different approach is used, the observed associations being presented in an order which seeks to reflect their chronological and causal relationships (see Section 13.1).

$$systolicBP = \beta_0 + \beta_1 dampness + \beta_2 sex + \beta_3 smoking + \varepsilon$$

Prior to performing regression modelling, it was necessary to assess whether the distributions of the three physiological variables satisfied the assumption of Normality. Initially, this was done by obtaining the Shapiro-Wilk *W* statistic (Shapiro & Wilk, 1965) for each variable; the associated *p* values were as follows (small *p* values indicate rejection of the null hypothesis of Normality):-

systolic blood pressure	-	$p = 0.003$
diastolic blood pressure	-	$p < 0.001$
standardised FEV_1	-	$p < 0.001$

However, these results must be interpreted with caution because the Shapiro-Wilk test is highly sensitive to even modest departures from Normality: "…in large samples the test is able to detect small amounts of non-Normality, that in most circumstances would be unimportant." (Altman 1991, p. 139). Therefore, the above results were supplemented by further investigation of the distributions. One frequently-used approach to assessing Normality (or conformance with other theoretical distributions) is to examine quantile-quantile ('Q-Q') plots of the variables. Such a plot "…graphs quantiles of the observed data against similar quantiles of a probability distribution conjectured to be a reasonable match. For a good fit, a Q–Q plot is roughly linear, with systematic deviations suggesting a lack of fit." (Sarkar 2008, p. 40). Normal Q-Q plots of the three physiological variables were examined, and it was concluded that there was convincing evidence of a material departure from Normality, most notably in the case of the FEV_1 outcome.

Some authorities suggest that, given an adequate sample size, such departures from Normality are nonproblematic in the context of regression analysis because the operation of the central limit theorem means that the assumption of Normality may be relaxed. A typical expression of this view is given by Katz:-

> "Multiple linear regression assumes a normal distribution... ...if your sample size is large (greater than 100), you can assume that the assumption of normal distribution is met." (Katz 2002, pp. 51-53)

While initially reassuring, the above statement is something of a simplification. The central limit theorem applies only to the point estimates of the means in a linear model, not to the

standard errors associated with those estimates. In the present context, this meant that while a regression model would arguably provide an acceptable estimate of the parameter of interest (e.g. the effect of a unit increase in accumulated exposure to dampness on systolic BP [coefficient β_1 in the notation used above]), confidence intervals directly derived from such a model (via the asymptotic standard error) would be unreliable. Recognising this, confidence intervals were estimated in these analyses via bootstrapping (Efron, 1979; Efron & Tibshirani, 1986), using the `boot` library in R software Version 2.7.1.

9.4 Analytical regime 2: logistic regression

A second group of relationships were assessed via logistic regression, the associations involved being:-

i. accumulated disadvantage *with* the clinical variables
ii. accumulated exposure to residential hazards *with* the clinical variables
iii. accumulated disadvantage *with* the medication usage variables
iv. accumulated exposure to residential hazards *with* the medication usage variables

These associations (all involving binary outcome measures) were assessed via a series of logistic regression models in which the presence of each outcome (e.g. heart disease, or bronchodilator use) was predicted by values of each cumulative measure, together with subjects' sex and smoking status. For example, the model investigating the relationship between accumulated disadvantage and heart disease took the form

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 disadvantage + \beta_2 sex + \beta_3 smoking\ ,$$

where p is the probability of a subject reporting heart disease.

Associations involving cumulative exposures - to both social disadvantage and residential risk - and the secondary health outcomes (those not expressing aspects of cardio-respiratory health) were also estimated using the logistic regression method. However, as explained earlier, results for these investigations are not presented in detail in the *Results*.

9.5 Analytical regime 3: initial investigation via Kruskal-Wallis test, conditionally followed by multiple regression

A third analytical approach was used to assess the following associations:-

i. time-dependent social location *with* the physiological variables
ii. time-dependent exposure to residential hazards *with* the physiological variables

Relationships between the typologies of time-dependent experiences (See Sections 6.3, 7.4 and 7.7) and the physiological variables were initially assessed via Kruskal-Wallis tests. Nine such tests were conducted in total (three classification schemes with three physiological variables). Although the K-W test is sometimes described as a test for equality of medians (e.g. Altman, 1991) or identity of distributions (e.g. Rice, 1995), its developers actually expressed its function in subtly different terms:-

> "…what *H* [the K-W test statistic] really tests is a tendency for observations in at least one of the populations to be larger (or smaller) than all the observations together, when paired randomly." (Kruskal & Wallis 1952, p. 598)

Thus, in the present context the K-W test was employed to determine whether values of the physiological parameter of interest (e.g. systolic blood pressure) were higher or lower in individuals characterised by specific patterns of experience (of, say, exposure to dampness) over the adult lifecourse, than in the sample as a whole.

In performing the above tests, the intention was that where a significant between-group difference was detected (at the conventional 5% level), the specific relationship between the time-dependent representation and the physiological measure involved would be examined in greater detail via multiple regression. For the investigations involving social position and exposure to air pollution, all K-W tests returned clearly nonsignificant results, so no further investigation was carried out. For the remaining time-dependent measure (exposure to dampness) no significant association with either diastolic blood pressure or standardised FEV1 was indicated by the K-W tests. However, a significant relationship with systolic BP was indicated, the test returning $p = 0.05$. Consequently, the relationship between subjects' clustered dampness experience and this outcome was investigated further via multiple regression. A model was fitted in which systolic BP was predicted by pollution group membership, sex and smoking status (current / former smoker *vs.* 'never smoker').

9.6 Analytical regime 4: Freeman-Halton test

A fourth group of associations was assessed via Freeman-Halton tests (Freeman & Halton, 1951), the relationships involved being:-

i. time-dependent social location *with* the clinical variables
ii. time-dependent exposure to residential hazards *with* the clinical variables
iii. time-dependent social location *with* the medication usage variables
iv. time-dependent exposure to residential hazards with the medication usage variables

The F-H test is a generalisation of Fisher's exact test to contingency tables with dimensions larger than 2 X 2. In the present context, the test assesses the probability of a general association between a clustering scheme representing (say) social position and a clinical outcome (e.g. the presence of heart disease). The analytical scenario may be visualised thus:-

heart disease present?	**social position cluster 1**	**social position cluster 2**	**social position cluster 3**	**social position cluster *n***
NO	*n1*	*n3*	*n5*	*etc.*
YES	*n2*	*n4*	*n6*	*etc.*

The F-H test was selected in preference to the chi-square test because the cross-classifications involved contained undesirably high numbers of cells with small expected frequencies. Under these conditions, chi-square testing would be in violation of the accepted 'rule of thumb' which states that 80% of the cells in the contingency table should have expected frequencies greater than 5, and all cells should have expected frequencies greater than 1 (Altman, 1991). The F-H test is not subject to this restriction: "The method... ...is generally of use in cases where χ^2 would be used were not certain of the observed and expected numbers too small." (Freeman & Halton 1951, p.149). For the analyses conducted using this approach, F-H tests were actually performed via Monte Carlo simulation because the time required to compute true F-H tests would be prohibitive.

Relationships between subjects' time-dependent experiences (of both social location and exposure to residential hazards) and the ten secondary health outcomes were also

investigated using Freeman-Halton tests. The results of these tests are not presented in detail in the *Results*.

9.7 Analytical regime 5: multiple regression

Associations between accumulated disadvantage and the measures of accumulated exposure to residential hazards were examined via multiple regression. A series of three models was constructed in which each of the exposure variables (damp, air pollution and total load) was predicted by (i) accumulated disadvantage, and (ii) the gender of the subject. Inclusion of the latter was justified on the grounds that exploratory analysis identified a significant between-gender difference in values of the disadvantage measure. The models fitted were of the form

$$hazard = \beta_0 + \beta_1 disadvantage + \beta_2 sex + \varepsilon,$$

the parameter β_1 providing an estimate of the change in the outcome (e.g. the number of years of exposure to residential damp) associated with a unit change in the predictor i.e. each additional year spent under conditions of social disadvantage. Because the distributions of all three residential hazard measures were heavily skewed, confidence intervals around the parameter estimates for β_1 were obtained via bootstrapping.

9.8 Analytical regime 6: initial investigation via Freeman-Halton test, followed by multinomial logistic regression

A sixth analytical regime served to investigate associations between the time-dependent representation of social position and the two measures of time-dependent housing conditions (i.e. dampness and air pollution). These relationships were initially assessed via Freeman-Halton tests; the analytical scenario may be visualised in the following form:-

	dampness cluster 1	**dampness cluster 2**	**dampness cluster 3**		**dampness cluster *n***
social position cluster 1	*n1*	*n4*	*n7*	...	*etc.*
social position cluster 2	*n2*	*n5*	*n8*	...	*etc.*
social position cluster 3	*n3*	*n6*	*n9*	...	*etc.*
	...	...	...	...	
social position cluster *n*	*etc.*	*etc.*	*etc.*	...	*etc.*

These tests were applied to determine the probability of a general association between the row variable (the clustered representation of social location) and the column variable (the grouped measure of exposure to dampness or to air pollution).

In planning the study, the original intention was that these particular associations would be explored in greater depth via multinomial logistic regression models (McFadden, 1974; Hosmer & Lemeshow, 2000). This class of model extends the binomial logistic regression model to cater for dependent variables with *n* unordered levels ($n > 2$), and could potentially provide information about specific patterns of association between grouped trajectories of social position and subjects' experiences over time of exposure to dampness and to air pollution. However, it was not feasible to fit such models due to the widespread presence of sparseness in the data. The problem is illustrated in Table 9.8.1 *(next page)*, which shows a crosstabulation of social position group (rows) with air pollution group (columns). In Table 9.8.1, almost half of the cells (34 of 72; = 47.2%) are unpopulated. Under conditions of such widespread sparseness, maximum likelihood estimates are not available (see Sections 6.3.1 and 6.3.7). Consequently, the planned use of multinomial logistic regression could not be implemented.

TABLE 9.8.1: Contingency table showing time-dependent classifications by socioeconomic position (rows) and air pollution exposure (columns). Cell content is number of subjects (upper element) and percentage of overall total (lower element).

Socioeconomic position cluster

	Air pollution cluster								
Frequency Percent	1	2	3	4	5	6	7	8	Total
1	44 20.75	11 5.19	7 3.30	1 0.47	2 0.94	0 0.00	2 0.94	0 0.00	67 31.60
2	18 8.49	0 0.00	7 3.30	3 1.42	7 3.30	1 0.47	1 0.47	0 0.00	37 17.45
3	28 13.21	1 0.47	7 3.30	3 1.42	4 1.89	1 0.47	0 0.00	1 0.47	45 21.23
4	16 7.55	3 1.42	6 2.83	2 0.94	2 0.94	1 0.47	0 0.00	2 0.94	32 15.09
5	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
6	7 3.30	2 0.94	3 1.42	1 0.47	0 0.00	0 0.00	0 0.00	0 0.00	13 6.13
7	7 3.30	0 0.00	2 0.94	1 0.47	0 0.00	0 0.00	0 0.00	1 0.47	11 5.19
8	3 1.42	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	3 1.42
9	2 0.94	1 0.47	0 0.00	1 0.47	0 0.00	0 0.00	0 0.00	0 0.00	4 1.89
Total	125 58.96	18 8.49	32 15.09	12 5.66	15 7.08	3 1.42	3 1.42	4 1.89	212 100.00

Frequency Missing = 82

***

This completes description of the methods used to assess the main associations examined in the study. However, as explained previously, subjects' detailed trajectories of experience over time (which were originally created as an interim stage in the construction of the variables used in the analyses described above) came increasingly to be seen as outcomes of interest in their own right. The following chapter outlines a further set of methods used to explore features of these sets of sequences.

CHAPTER 10: METHODS (VI) – EXPLORING PATTERNS OF EXPERIENCE OVER TIME

10.1 A graphical representation of social location: the Position Weight Matrix

The detailed sequences or trajectories of experience described earlier contain considerable information about subjects' lives. However, it can be difficult to discern features or patterns of interest from examination of the raw sequences *en masse*. Consider the trajectories of year-on-year social location, which are shown in Appendix 5. Questions such as whether the prevalence of (say) the non-manual state tends to increase or decrease with age, or whether local peaks of the non-manual condition occur at specific ages, cannot readily be answered from the raw information. This section outlines a method developed to explore features of the social position sequence data, with the goal of identifying pattern and structure which may not be evident from simple visual scrutiny of the entire set of social trajectories.

One useful device for characterising the properties of a family of sequences is the Position Weight Matrix (PWM), which is used in the field of sequence data mining (Dong & Pei, 2007). The concept may be illustrated using a small subset of data from the study. In the interests of simplicity, this initial illustration uses data drawn from the trajectories of dampness exposure, in which only two states are permitted (i.e. exposed or unexposed). Figure 10.1.1 *(next page)* shows a segment of six detailed sequences of dampness exposure. This segment (hereafter referred to in this section as a *window*, in conformance with the terminological conventions of Dong & Pei) covers the age range from 25 to 30 (six datum points). The figure uses the now-familiar notation 0 = not exposed to damp at age *Y* years, 1 = exposed at age *Y*. In this discussion, the set of permitted values (i.e. 0 and 1) will be referred to as the *alphabet* of the sequence[33]. For a sequence set which features an alphabet of *a* unique values (here, two [zero and one]) and a window of width *w* elements (here, six), the PWM is an *a* X *w* matrix showing the proportion of occasions on which each possible value in the alphabet occurs at each position in the window. To illustrate, Figure 10.1.2 *(next page)* shows a 2 X 6 PWM for the data of Figure 10.1.1. The top line of the figure shows that alphabet value zero (unexposed) was found for 4 of the 6 sequences at age 25; for 3 of 6 at age 26; for 4 of 6 at age 27; and so on. Similarly, alphabet value one (exposed) was recorded for 2 of 6 sequences at age 25; for 3 of 6 sequences at age 26 *etc*.

[33] Use of the term 'alphabet' to denote the range of possible values for the elements in a sequence is widespread in the sequence comparison literature (e.g. Kruskal, 1999; Erickson & Sellers, 1999).

FIGURE 10.1.1: A window of width six (ages 25 to 30 years) extracted from six sequences of dampness exposure. Explanation of symbols appears beneath figure.

	AGE					
SEQUENCE NO.	25	26	27	28	29	30
1	...0	1	1	1	0	0...
2	...1	1	0	0	0	0...
3	...1	1	1	0	0	0...
4	...0	0	0	0	0	0...
5	...0	0	0	0	0	1...
6	...0	0	0	0	1	1...

NOTE: Symbols identify dampness status thus: '0' = not exposed to damp, '1' = exposed to damp.

FIGURE 10.1.2: Position weight matrix for the data of Figure 10.1.1.

	AGE					
ALPHABET VALUE	25	26	27	28	29	30
0 (not exposed)	4/6	3/6	4/6	5/6	5/6	4/6
1 (exposed)	2/6	3/6	2/6	1/6	1/6	2/6

From the PWM in Figure 10.1.2, it is possible to ascertain that exposure was most frequently encountered at age 26, and least frequently experienced at ages 28 and 29. Of course, in this trivial illustration these features could instantly be grasped from the raw data of Figure 10.1.1. However, where the number of sequences is greater, and the window width and alphabet more extended than in this trivial example, the PWM is a potentially useful tool for detecting patterns and trends in the sequence set. This utility is demonstrated in Figure 10.1.3 *(next page)*, which shows a PWM representing subjects' social location at the ages from 15 to 25 years[34]; the values shown are the proportions of the total number of subjects (*n* = 294) holding each social state at each age point. Each column (age point) sums to one (i.e. 100% of all cases) after allowing for rounding errors.

[34] This range may seem arbitrary - in fact, it was chosen as the maximum window (from the temporal start of the sequence i.e. age 15) which could comfortably be accommodated in the physical page width while using a font of legible size.

FIGURE 10.1.3: Position weight matrix showing socioeconomic position for a sequence window covering the age range 15 to 25 years. Explanation of symbols in the 'STATE' column appears beneath figure. Values shown are the proportion of subjects in social state *X* at age *Y* years.

	AGE										
STATE	15	16	17	18	19	20	21	22	23	24	25
MAN	0.77	0.77	0.76	0.52	0.40	0.45	0.49	0.51	0.53	0.56	0.57 ...
NM	0.23	0.23	0.23	0.26	0.25	0.29	0.32	0.33	0.34	0.35	0.37 ...
AF	0.00	0.00	0.00	0.10	0.11	0.09	0.07	0.06	0.05	0.04	0.03 ...
NEMP	0.00	0.00	0.00	0.13	0.24	0.17	0.12	0.10	0.07	0.05	0.03 ...

NOTE: The 'STATE' column identifies subjects' social location thus: 'MAN' = manual, 'NM' = non-manual', 'AF' = Armed Forces, 'NEMP' = non-employed.

A number of features of interest are evident in the figure. The proportion of individuals holding manual status (top line) broadly declines with increasing age, while the corresponding proportion in the non-manual state (second line) increases with age. Considered jointly, these suggest an effect of upward social mobility in this age range, though the trend is possibly obscured to some extent by the influence of Armed Forces service. Military service itself (third line) exhibits a local peak in the age range from 18 to 20, declining thereafter. This is of course entirely expected, reflecting the reality that soldiering is, in the main, a profession for young men. The final line (representing non-employment) also shows a local peak, this time in the age range 18-22. This may reflect the withdrawal of young women from the labour market for the purposes of child-bearing. The generation of separate PWMs by sex would provide further indications relating to this hypothesis, but this is not pursued at this stage. The main purpose of Figure 10.1.3 is to demonstrate the potential of the PWM as a descriptive device for revealing patterns and trends in sequence data.

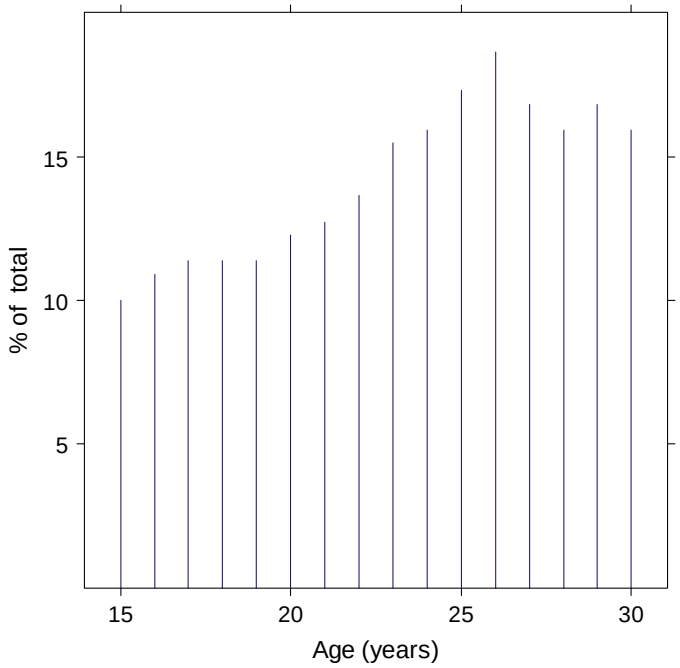
Where the sequences examined are lengthy (that is, consist of a large number of individual elements), the PWM in the form discussed above becomes of limited value because it is not possible legibly to display very long strings of numerics on a physical page or data display device. However, the concept of the PWM may be translated directly into a graphical form, in which the numerical proportions of Figure 10.1.3 are encoded via a colour or greyshade gradient. For example, proportions between zero and (say) < 0.1 might be represented by colour *A*, values from 0.1 to < 0.2 by colour *B*, and so on. Such a graphical position weight matrix (GWPM) was prepared to illustrate the set of trajectories of social location. The result is given as Figure 14.2.1 in *Results* Section 14.2.

10.2 A graphical representation of residential hazard exposure - the 'spike' chart

While the GWPM is a useful device for exploring a set of sequences in which each element can assume one of a number of possible values, it is arguably less appropriate to represent sequences in which only two possible states are recognised (e.g. exposed or not exposed to dampness). For data of this kind, a GPWM will consist of only two rows: one showing the proportion of subjects with the characteristic of interest, the other representing the proportion without. In this situation, the relationship between the rows at specific points (the illustration of which is one of the main strengths of the GPWM) becomes entirely predictable. If the value of one row at any point is the proportion *p*, the value of the other row at the same point is always 1 - *p*. Because of this, the GPWM offers no particular advantage over a number of other charting techniques for sequences of binary data. In fact, the graphical position weight matrix may actually be inferior to alternative approaches. A limitation of the GPWM is that its resolution is constrained by the number of colours which comprise the colour gradient. Depending on the colour encoding scheme used, a small change in proportion (say, from 0.11 at age *Y* years to 0.14 at age *Y*+1) might not be visible, because both proportions could fall within the range represented by a single colour value. The obvious solution - increasing the number of colours in the gradient - is actually of limited value in resolving this problem, because the restricted ability of the human eye to detect subtle colour variations in small areas (whether on the printed page or the computer display screen) limits the number of distinct colours which may be used. Alternative graphical methods which display proportions as (say) line segments of varying length actually offer a level of visual resolution which is superior to the relatively coarse colour encoding scheme of the GPWM.

One useful alternative to the graphical PWM for exploring time-related trends in sequences of binary elements is the 'spike' chart (essentially a simple barchart with zero-width bars). An example is given in Figure 10.2.1 *(next page)*, which shows the proportion of the total number of individuals in the dataset who were exposed to damp at each individual yearly age point between the ages of 15 and 30. The figure displays essentially the same information as would be shown in a GPWM (that is, the proportion of 'positives' at each point), but arguably does so with greater precision. A set of such spike charts was created to explore the sequences of exposure to dampness and to air pollution; these are presented in Section 14.3.

FIGURE 10.2.1: Proportion (percentage) of individuals recording exposure to dampness at each year in the age range from 15 to 30.



10.3 Further exploration of the social location sequences: methods based on *k*-grams

Compared to the trajectories of exposure to residential hazards, the social position sequences are characterised by greater complexity, in the sense that four distinct states are recognised for each age point in the latter (as against two in the former). Because of this, techniques additional to that of Section 10.1 may beneficially be employed to gain a richer insight into how respondents' social location varies over time. One approach to detecting interesting structure in the sequences of social position involves the use of *k*-grams[35]. (Dong & Pei, 2007). *K*-grams are substrings of length *k* elements, defined more formally thus: "Let $k \geq 1$ be a positive integer. A *k*-gram is a sequence over [an alphabet] $\boldsymbol{A}$ of length *k*." (Dong & Pei 2007, p.49).

To illustrate the concept, Table 10.3.1 *(next page)* shows the 12 possible 2-grams (i.e. *k*-grams of length 2 elements) representing the various forms of social transition between age *Y* years and age *Y*+1 which may be present in the sequences of social position (see Section 6.1). Examining frequency tables of, or summary values for, these 2-grams (either

[35] Like other terms with a leading variable, *k*–grams are often referred to by other designations, such as '*n*-grams' (see for example Venkataraman, 2001).

individually or in groups) can extract potentially useful information about subjects' experiences over time. Unsuspected features of the data may be revealed, as is demonstrated by the following simple example.

TABLE 10.3.1: Listing of possible 2-grams which indicate a shift in socioeconomic position.

2-gram (0 = manual, 1 = non-manual, * = Armed Forces, . = non-employed)	**type of status change**
01	manual ➔ non-manual
0.	manual ➔ non-employed
0*	manual ➔ Armed Forces
10	non-manual ➔ manual
1.	non-manual ➔ non-employed
1*	non-manual ➔ Armed Forces
.0	non-employed ➔ manual
.1	non-employed ➔ non-manual
.*	non-employed ➔ Armed Forces
*0	Armed Forces ➔ manual
*1	Armed Forces ➔ non-manual
*.	Armed Forces ➔ non-employed

Among the 294 subjects in the dataset, the average number of instances of the 2-gram '01' (which indicates a transition from the manual to the non-manual state) is 0.68. The corresponding mean value for the number of instances of the 2-gram '10' (indicating a non-manual to manual transition) is 0.47. However, these overall summary values mask a substantial between-gender difference. For males, the respective average numbers of occurrences of these 2-grams are 0.47 (for 2-gram '01'; upward mobility) and 0.27 (for 2-gram '10'; downward mobility). Among female subjects, the corresponding means are 0.87 and 0.65. It therefore appears that both of these specific types of social transition are markedly more common among women than among men in the sample[36]. This feature of the social position sequences cannot easily be discerned by examining the raw trajectories.

K-grams were used to explore three aspects of the sequences of social position. First, an investigation was performed to determine the frequency with which each of the 12 types of state transition shown in Table 10.3.1 was observed in the data. The objective was to identify the types of social transition event which respondents experienced most often. A particular focus of interest was whether the classic configurations of upward and downward

[36] In fact, interpretation of this particular finding must involve caution. The apparently higher numbers of these specific transitions among females may be partly artefactual, reflecting the imputation to married women of their husbands' social status under certain circumstances (see Section 6.1.3). This possibility is not pursued here, as the example is introduced purely to illustrate a potential use of the *k*-gram concept.

mobility (represented respectively by shifts from manual to non-manual, and from non-manual to manual status) were more or less common than the other state changes shown in Table 10.3.1. Second, *k*-grams were used to determine the frequency with which respondents' social status changed at each age point. This investigation sought to establish whether changes in social status were concentrated at certain ages. Finally, *k*-grams were used to examine the distribution of the total number of transitions experienced by subjects. The aim in this case was to provide a broad determination of the extent to which respondents' lives were characterised by social flux (or, alternatively, by social stasis). The results of these investigations appear in Section 14.4.

10.4 Representing joint experience of multiple hazards: a method based on 3-tuples

A limitation of the methods described above is that they reveal little about trends over time in respondents' joint exposure to multiple hazards. In order to investigate such trends, it is necessary to construct a representation of the subject's concurrent experience over time of all three factors of interest. Creation of such a representation is now described, with reference to Figure 10.4.1. This shows one subject's exposure to the three hazards examined in the study at six individual age points. From this, it is possible to determine that at the age of 18 this respondent experienced both social disadvantage and air pollution, but was not subject to dampness. This pattern of joint experience persisted until age 21. However, at 22 years the person encountered exposure to dampness but ceased to be subject to air pollution, this combination of circumstances being preserved into age 23.

FIGURE 10.4.1: Observed hazard exposure for subject P132 at six consecutive age points.

	WHETHER EXPOSED TO HAZARD AT AGE...					
HAZARD	18	19	20	21	22	23
Social disadvantage	YES	YES	YES	YES	YES	YES
Dampness	NO	NO	NO	NO	YES	YES
Air pollution	YES	YES	YES	YES	NO	NO

The subject's status with respect to these three hazards at age *Y* years may be expressed symbolically using the data structure known as a *tuple* (Dong & Pei, 2007). Specifically, the information presented in Figure 10.4.1 may be completely represented by a sequence of 3-tuples (i.e. tuples consisting of three components) of the general form

$$(Ax, By, Cz),$$

where A, B and C indicate (respectively) the three risk factors, and *x*, *y* and *z* the person's status (exposed or free from exposure) to each factor at that age point. Adopting a more informative notation, the status of respondent P132 (the individual featured in Figure 10.4.1) at age 18 may be represented as

$$(\text{DISA+}, \text{DAMP-}, \text{POLL+}),$$

where the labels DISA, DAMP and POLL provide concise but meaningful identifiers for the three risk factors, and + and - are status flags indicating respectively that the subject was or was not exposed to each hazard. Thus, the information of Figure 10.4.1 may be expressed as a sequence of 3-tuples:-

```
START (age 18)     (DISA+,DAMP-,POLL+),
                   (DISA+,DAMP-,POLL+),
                   (DISA+,DAMP-,POLL+),
                   (DISA+,DAMP-,POLL+),
                   (DISA+,DAMP+,POLL-),
                   (DISA+,DAMP+,POLL-)    END (age 23)
```

Using the 3-tuple structure, it is possible to create a graphical position weight matrix (see Section 10.1) which represents the proportion of the total number of subjects who experienced each *combination* of exposures at each age point. The general form of such a presentation is shown in Figure 10.4.2 *(next page)*. A complete graphical position weight matrix corresponding to Figure 10.4.2, in which the proportions p_{nn} are encoded via a colour gradient, is presented in Section 14.5 (Figure 14.5.1a / b).

FIGURE 10.4.2: General form of a position weight matrix showing the proportion of subjects exposed to each combination of hazards at each age point. Hazard combinations (rows) are expressed as 3-tuples representing exposure (symbol '+') or freedom from exposure (symbol '-') to each hazard type (DISAdvantage, DAMPness and POLLution). Proportions sum to 1 (or 100%) columnwise.

	AGE					
JOINT EXPOSURE STATUS (expressed as 3-tuple)	N	N+1	N+2	N+3	N+4	etc.
(DISA+,DAMP+,POLL+)	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	..
(DISA+,DAMP+,POLL-)	p_{21}	p_{22}	p_{23}	p_{24}	p_{25}	..
(DISA+,DAMP-,POLL-)	p_{31}	p_{32}	p_{33}	p_{34}	p_{35}	..
(DISA+,DAMP-,POLL+)	..					
(DISA-,DAMP-,POLL-)	..					
(DISA-,DAMP-,POLL+)	..					
(DISA-,DAMP+,POLL+)	..					
(DISA-,DAMP+,POLL-)	..					

***

With this, explanation of the analytical methods used in the study is complete; the presentation of results now begins.

CHAPTER 11: RESULTS (I) - PROPERTIES OF THE DERIVED MEASURES OF SOCIAL POSITION AND RESIDENTIAL CONDITIONS

11.1 Introduction and structure of results presentation

This chapter presents descriptive material (and, where appropriate, summary statistics) for the measures of social location and residential conditions which were used in the study. These measures have been described in the relevant *Methods* sections, and are of two types. The first type consists of quasi-continuous numeric variables (more precisely, integer quantities) representing accumulated levels of exposure, expressed in years. With one exception (see Section 7.8) these measures are constrained to fall within the range of integers from zero (indicating no exposure to the factor of interest over the age range 15 to 60 years) to 46 (representing continuous exposure across the period). The second type of measure comprises categorical classification schemes, constructed via the application of optimal matching and cluster analysis to the original detailed sequences. These schemes group subjects' time-ordered experiences of the factor of interest (e.g. exposure to air pollution) into relatively small numbers of higher-level classes which ideally represent broadly similar patterns or trajectories of experience.

For each numeric (cumulative) variable, the chapter presents (i) a graphical representation of the variable's distribution, and (ii) a table of summary statistics. For each of the categorical (clustered) measures, the present chapter is restricted to providing a pointer to an Appendix which presents the group structure in detail. The cluster schemes cannot conveniently be included 'in stream' with the chapter text due to excessive length. The structure of the chapter is summarised in Table 11.1.1.

TABLE 11.1.1: Structure of Chapter 11

section	**measure described**	**related section(s) in *Methods* chapters**
11.2	accumulated social disadvantage	6.2
11.3	classification scheme for social position	6.3
11.4	accumulated exposure to residential dampness	7.3
11.5	classification scheme for exposure to residential dampness	7.4
11.6	accumulated exposure to air pollution	7.6
11.7	classification scheme for exposure to air pollution	7.7
11.8	accumulated exposure to damp or air pollution	7.8

The present chapter is followed by four others which present additional results. Summary statistics and descriptive material covering the main health outcome measures (see Section

5.3.3) are given in Chapter 12. Chapter 13 reports findings relating to the main associations investigated by the study, while Chapter 14 examines properties of the original sequences of experience (prior to classification); derivation of these sequences is described in Sections 6.1, 7.2 and 7.5. Finally, material relating to the indicators of occupational exposure (see Chapter 8) is presented in Chapter 15. A principle strictly observed throughout the *Results* chapters is that no discussion or comment is made on the reported findings; this material is intended only to present the results. All interpretation is restricted to the *Discussion* chapters.

11.2 Accumulated social disadvantage

The distribution of the measure of accumulated disadvantage is shown graphically in Figure 11.2.1, and summary statistics for this quantity are given in Table 11.2.1 *(next page)*. As explained earlier in the relevant *Methods* section, values of this measure were calculated for 285 of the 294 cases in the dataset. The figure and table show results separately for each sex; the rationale behind this approach (which implies recognition of a possible between-sex difference in the experience of disadvantage) is now described.

FIGURE 11.2.1: Distribution of accumulated social disadvantage over the age range 15-60 years (subjects classified by sex). Sample numbers are n = 133 (male); n = 152 (female).

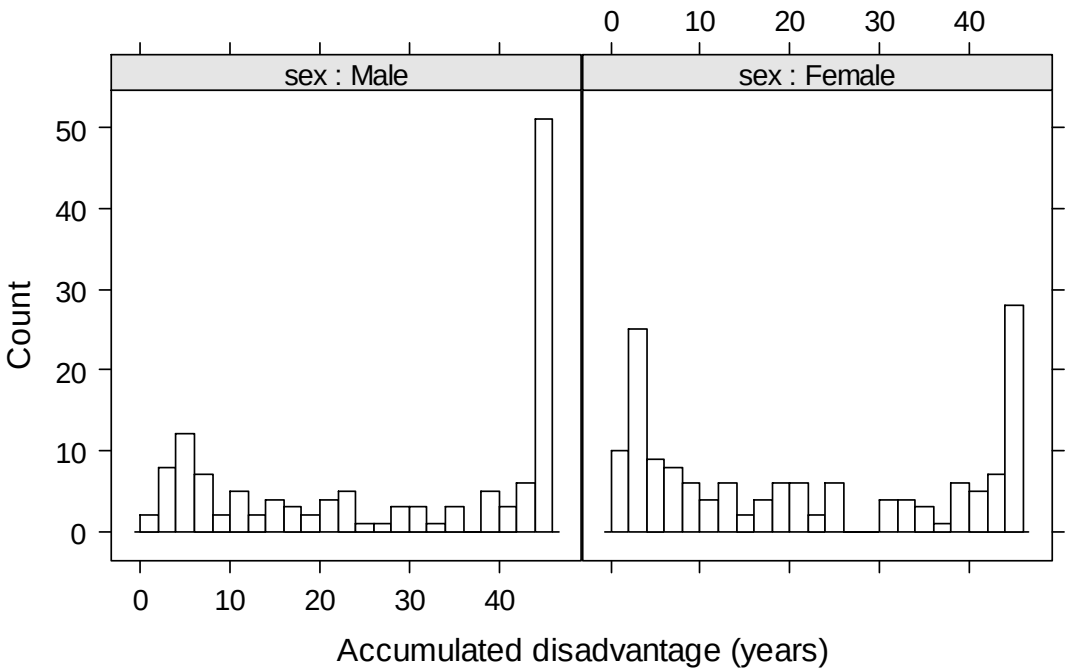


TABLE 11.2.1: Summary statistics for accumulated social disadvantage (values in years).

	males (n = 133)	**females (n = 152)**	**all (n = 285)**
mean (SD)	29.4 (17.0)	22.2 (17.1)	25.5 (17.4)
median	35.0	19.0	24.0
minimum	2	0	0
maximum	46	46	46

NOTE: Mann-Whitney test for equality of values between sexes returned $p < 0.01$.

Under the conceptual model examined in this study, socioeconomic position (more specifically, social disadvantage) is a putative risk factor for exposure to poor housing conditions, and – via a hypothesised mediating influence of such conditions - for cardio-respiratory disease (see Figure 4.2.1 in Chapter 4). As in many epidemiological analyses, the possibility that a risk factor (here, disadvantage) may vary among sub-groups of the population of interest required that the extent of such variation be assessed and, if necessary, adjusted for in analysis. In the specific context of this study, there were plausible reasons to hypothesise that accumulated disadvantage might exhibit some systematic variation by sex. The uncertainties associated with determining women's SEP via occupation-based measures have been discussed in Section 6.1.3, which outlined how, in the special case of married women, the social location presumed to apply was in certain circumstances that of the woman's husband. It was recognised that the non-uniform approaches used to determine SEP (and hence to quantify disadvantage) for men and for women may have led to differences in the respective values observed for each sex. While such quasi-artefactual variation cannot be distinguished from any 'true' sex-based differences in the experience of social disadvantage, it was considered appropriate to at least establish whether an appreciable between-sex difference in exposure to disadvantage was present, even though the interpretation of any such difference might be problematical. Some informal insight into between-sex variation is provided by the graphical presentation of Figure 11.2.1 and the summary values of Table 11.2.1. However, a more formal determination of whether cumulative disadvantage varied by sex was achieved by application of a Mann-Whitney test (see footnote to Table 11.2.1).

11.3 Time-dependent socioeconomic position

The clustered representation of time-dependent SEP which was adopted for use in analysis is shown in full in Appendix 2.

11.4 Accumulated exposure to residential dampness

Figure 11.4.1 shows the distribution of subjects' accumulated exposure to residential dampness. Summary statistics for this measure are presented in Table 11.4.1. Valid values are available for 220 of the 294 cases in the dataset. Following the approach adopted above for cumulative disadvantage, results are given separately for males and for females. In the case of disadvantage, the justification for presenting results separately by sex was fairly clear; however, the motivation for doing so with regard to the experience of dampness is more subtle, and is now outlined.

FIGURE 11.4.1: Distribution of accumulated exposure to residential dampness over the age range 15-60 years (subjects classified by sex). Sample numbers are n = 97 (male); n = 123 (female).

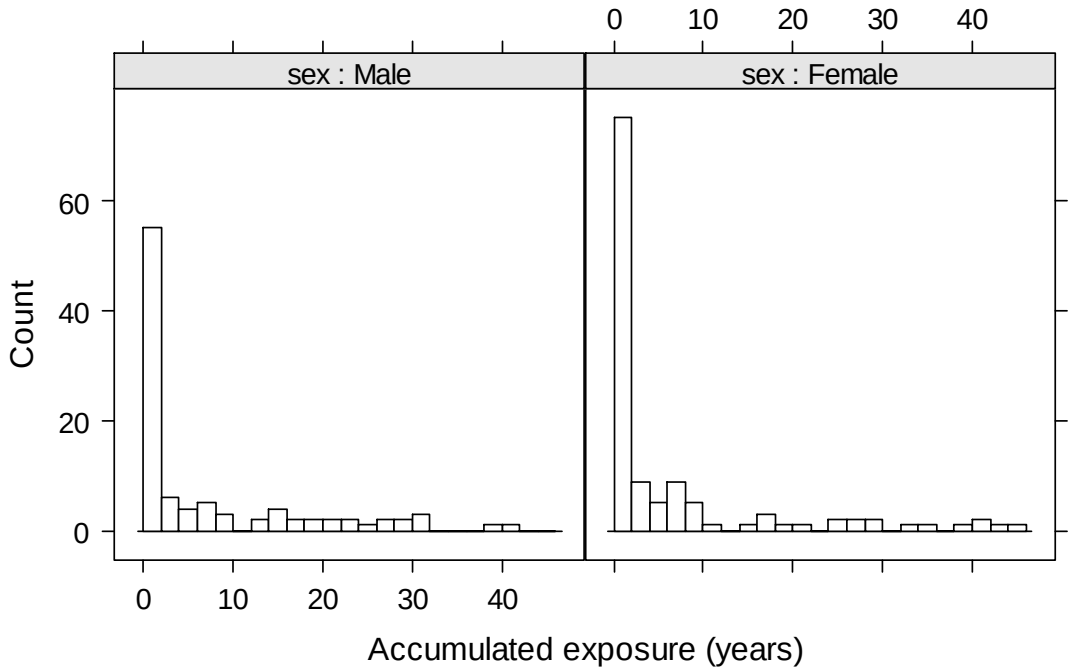


TABLE 11.4.1: Summary statistics for accumulated exposure to residential dampness (values in years).

	males (n = 97)	**females (n = 123)**	**all (n = 220)**
mean (SD)	7.0 (10.6)	6.2 (11.0)	6.5 (10.8)
median	0.0	0.0	0.0
minimum	0	0	0
maximum	42	46	46

NOTE: Mann-Whitney test for equality of values between sexes returned p = 0.64.

As explained in Section 5.3.2, subjects' exposure over the lifecourse to residential dampness was based solely on their recollections of the presence in the home of black mould and other manifestations of damp. While there is no *prima facie* reason to believe that women would

tend to live in damper homes than men (or *vice versa*), it is arguable that *perceptions* of residential dampness might vary between the sexes. There are various routes via which such sex-related differences in perception might arise. For example, it may be postulated that women in this sample would generally spend more time in the dwelling than men (due to the former acting as homemakers or full-time mothers), and thus have greater opportunity to become aware of the physical manifestations of dampness. In a variation on this hypothesis, it could be argued that certain distinct domestic roles fulfilled by women (specifically, tasks related to cleaning and upkeep of the home) might bring them into closer contact with mould and other signs of dampness than men. Because such plausible reasons for between-sex variation in the reported experience of dampness can be identified, it was considered appropriate to supplement the presentations of Figure 11.4.1 and Table 11.4.1 with a formal test of the null hypothesis that accumulated exposure to residential dampness does not vary between men and women (see footnote to Table 11.4.1).

11.5 Time-dependent exposure to residential dampness

Appendix 3 shows the clustered representation of time-dependent exposure to residential dampness.

11.6 Accumulated exposure to air pollution

The distribution of respondents' accumulated exposure to air pollution is illustrated in Figure 11.6.1 *(next page)*, summary statistics being given in Table 11.6.1 *(next page)*. Values of this quantity are available for 215 subjects. As for the two other accumulated measures reported above, exposure to air pollution is presented separately for men and for women. This is done largely in the interests of maintaining a consistent approach in the reporting of cumulative exposures, but is also partly motivated by the argument which is now outlined.

FIGURE 11.6.1: Distribution of accumulated exposure to air pollution over the age range 15-60 years (subjects classified by sex). Sample numbers are n = 96 (male); n = 119 (female).

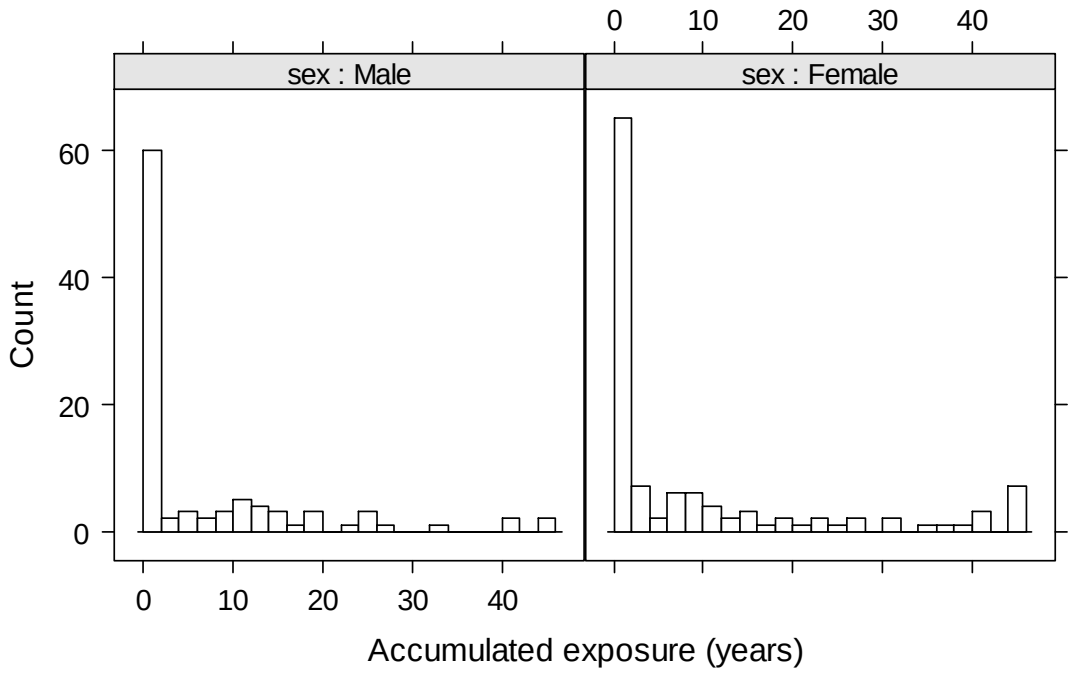


TABLE 11.6.1: Summary statistics for accumulated exposure to air pollution (values in years).

	males (n = 96)	**females (n = 119)**	**all (n = 215)**
mean (SD)	6.6 (11.1)	9.1 (14.1)	8.0 (12.9)
median	0.0	0.0	0.0
minimum	0	0	0
maximum	46	46	46

NOTE: Mann-Whitney test for equality of values between sexes returned p = 0.29.

As described in Section 5.3.2, assessment of subjects' exposure to air pollution was based on their recollections of the proximity of roads and industrial activity to the home. Although the possibility is rather tenuous, it may be hypothesised that men would – via their greater involvement in the world of regular paid employment – have a more specific awareness of the presence of factories and other industrial infrastructure in the environs of the home than women. For example, while a woman might be vaguely aware that some kind of commercial premises existed some streets away from her home, a man who actually worked in a local factory would fully appreciate the genuinely industrial nature of the activities performed there. Such differences of perception might result in some sex-related variation in subjects' reported exposure to air pollution. As stated, the possibility is rather remote; however, it was felt to be sufficiently plausible as to justify performing a formal test for between-sex variation in the experience of air pollution. The result of this test is presented as a footnote to Table 11.6.1.

11.7 Time-dependent exposure to air pollution

The grouped representation of time-dependent exposure to air pollution is shown in Appendix 4.

11.8 Accumulated exposure to residential dampness *or* air pollution: 'total hazard load'

One final accumulated measure represented the individual's total level of exposure (in years) to either dampness or air pollution (see Section 7.8). Unlike the other accumulated variables, this measure does not have a time-dependent (clustered) counterpart. The distribution of this quantity, which is referred to hereafter as the subject's total hazard load, is illustrated in Figure 11.8.1 *(next page)*. Note that due to the method of its calculation (summing the number of years' exposure to dampness and to air pollution) permissible values of this variable range from zero to 92. Values are available for 210 cases. Following the approach adopted throughout this chapter, results are presented separately for men and for women.

Table 11.8.1 shows summary statistics for this measure.

TABLE 11.8.1: Summary statistics for total hazard load (values in years).

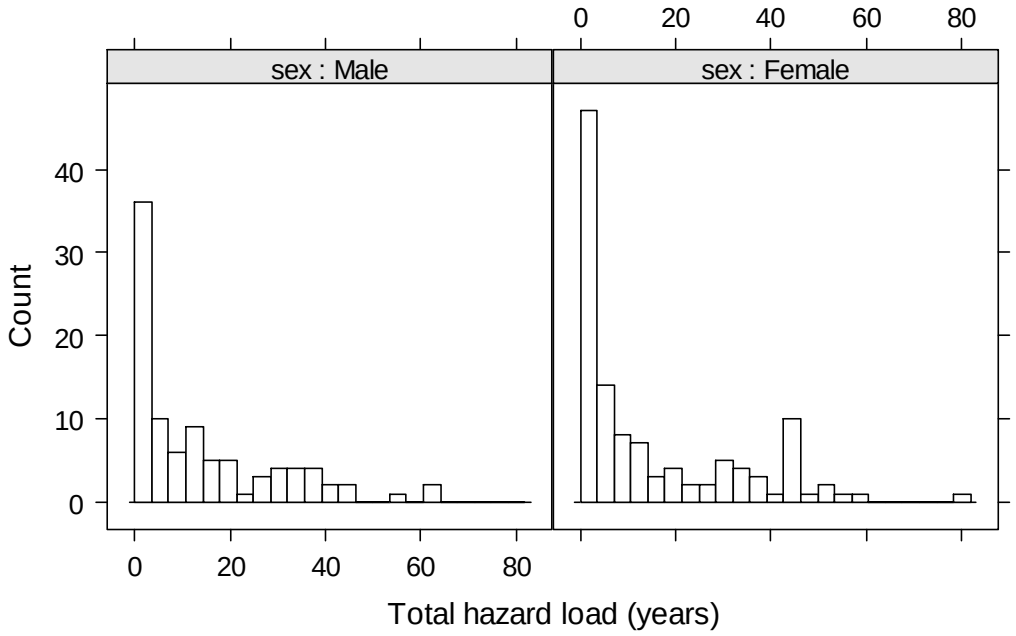
	males (n = 94)	**females (n = 116)**	**all (n = 210)**
mean (SD)	13.7 (15.6)	14.9 (18.1)	14.4 (17.0)
median	8.5	6.5	7.0
minimum	0	0	0
maximum	61	81	81

NOTE: Mann-Whitney test for equality of values between sexes returned $p = 0.90$.

***

The next chapter presents descriptive material relating to the health outcome measures (see Section 5.3.3).

FIGURE 11.8.1: Distribution of accumulated exposure to dampness or air pollution ('total hazard load') over the age range 15-60 years (subjects classified by sex). Sample numbers are *n* = 94 (male); *n* = 116 (female).



CHAPTER 12: RESULTS (II) - PROPERTIES OF THE HEALTH OUTCOME MEASURES

12.1 Introduction

Section 9.1 introduced a scheme for grouping the health outcome measures thematically, and this classification system is adhered to in the present chapter. Summary material is provided for those outcomes which relate most directly to the theme of the study; that is, those which express aspects of cardio-respiratory health. The variables which represent physiological measurements (blood pressure and lung function) are considered first, followed by the binary indicators denoting the presence or absence of specific disease types (heart disease, lung disease etc.). Finally, the dichotomous quantities representing medication use are discussed. This presentation reflects a chronological and causal ordering which links these measures: physiological change precedes clinical diagnosis of disease, which in turn results in medication use. This ordering is also observed when reporting on the associations observed in the study (Chapter 13).

As explained earlier, the secondary health outcomes (that is, the ten binary indicators which do not relate directly to cardiovascular or respiratory health - see Section 9.1) are not considered here.

12.2. Physiological variables

12.2.1 Systolic blood pressure

The distribution of systolic blood pressure in the sample is illustrated in Figure 12.2.1 *(next page)*. Because this variable is a genuinely continuous quantity (unlike the measures of accumulated exposure described in Chapter 11), the presentation takes the form of a kernel density estimate plot[37] rather than a histogram. Values are available for 291 subjects. In conformance with the approach used in the previous chapter, the distribution is shown separately for males and for females. Summary statistics for the variable are presented in Table 12.2.1 *(next page)*.

[37] Kernel density estimation (Silverman, 1986) involves estimating a probability density function by averaging across the observed data points to create a smooth approximation. A plot of the resulting values may be thought of as a 'smoothed histogram' (Hintze & Nelson, 1998).

FIGURE 12.2.1: Distribution of systolic blood pressure (subjects classified by sex). Sample numbers are n = 139 (male); n = 152 (female).

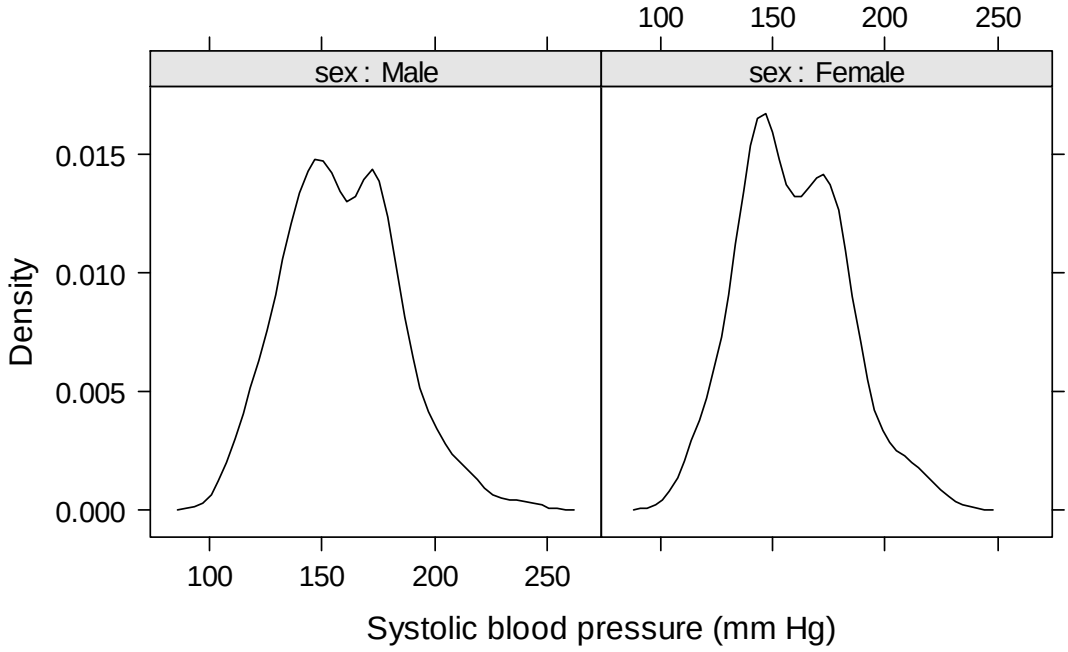


TABLE 12.2.1: Summary statistics for systolic blood pressure (values in mm Hg).

	males (n = 139)	**females (n = 152)**	**all (n = 291)**
mean (SD)	158.7 (24.4)	158.9 (23.1)	158.8 (23.7)
median	156.0	156.2	156.0
minimum	110.5	110.5	110.5
maximum	237.5	225.0	237.5

NOTE: Mann-Whitney test for equality of values between sexes returned p = 0.94.

12.2.2 Diastolic blood pressure

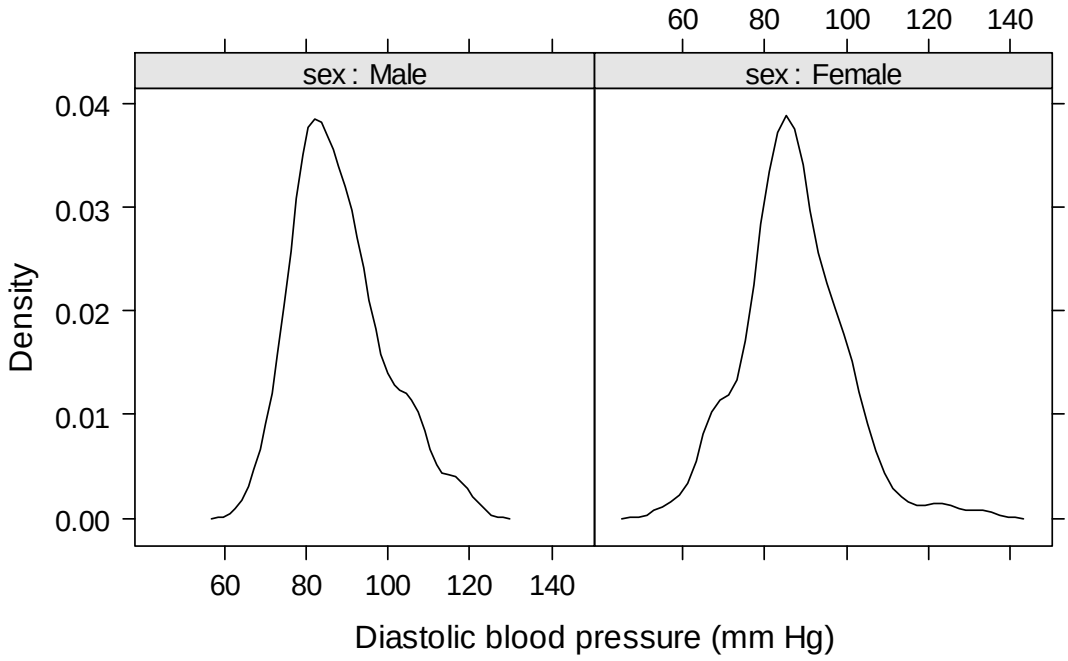
Figure 12.2.2 *(next page)* illustrates the distribution of diastolic blood pressure; valid values are available for 291 cases. Summary measures for this variable are presented in Table 12.2.2.

TABLE 12.2.2: Summary statistics for diastolic blood pressure (values in mm Hg).

	males (n = 139)	**females (n = 152)**	**all (n = 291)**
mean (SD)	88.2 (11.1)	86.9 (12.1)	87.6 (11.6)
median	86.5	85.8	86.0
minimum	67.5	56.0	56.0
maximum	119.0	132.5	132.5

NOTE: Mann-Whitney test for equality of values between sexes returned p = 0.60.

FIGURE 12.2.2: Distribution of diastolic blood pressure (subjects classified by sex). Sample numbers are *n* = 139 (male); *n* = 152 (female).



12.2.3 Standardised FEV_1

The distribution of standardised FEV_1 is shown in Figure 12.2.3; summary statistics are given in Table 12.2.3 *(next page)*. Values of this variable are available for 291 respondents.

FIGURE 12.2.3: Distribution of standardised FEV_1 (subjects classified by sex). Sample numbers are *n* = 137 (male); *n* = 154 (female).

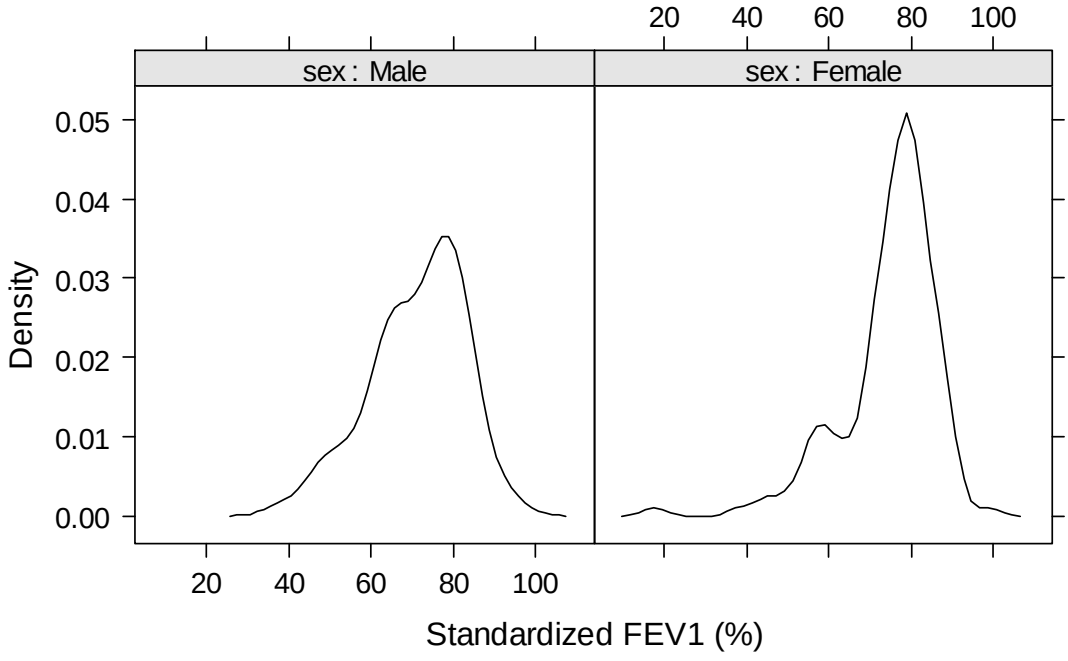


TABLE 12.2.3: Summary statistics for standardised FEV_1 (%).

	males (n = 137)	**females (n = 154)**	**all (n = 291)**
mean (SD)	71.2 (11.6)	74.7 (11.4)	73.1 (11.6)
median	73.0	77.5	75.4
minimum	37.3	17.3	17.3
maximum	95.7	98.8	98.8

NOTE: Mann-Whitney test for equality of values between sexes returned $p < 0.01$.

12.3. Clinical variables

Table 12.3.1 provides summary information (in the form of prevalence values) for the four binary health indicators which represent subjects' current (at the time of interview) experience of cardio-respiratory disease. For each outcome, the observed prevalence is shown separately for males and females. Equality of prevalence between sexes was assessed by performing a Fisher's exact test for each outcome.

TABLE 12.3.1: Observed prevalence values for four specific disease types. Cell content is number of positives (individuals reporting the condition), followed by % of positives, followed by total number of cases.

disease type	**MALES** ***n* positive / % prevalence (total *n*)**	**FEMALES** ***n* positive / % prevalence (total *n*)**	**ALL** ***n* positive / % prevalence (total *n*)**	***p***
heart disease	31 / 22.3 (139)	21 / 13.6 (155)	52 / 17.7 (294)	0.07
lung disease	14 / 10.1 (139)	17 / 11.0 (155)	31 / 10.5 (294)	0.85
stroke	5 / 3.6 (139)	7 / 4.5 (155)	12 / 4.1 (294)	0.77
high blood pressure	33 / 23.7 (139)	40 / 25.8 (155)	73 (24.8) / 294	0.69

NOTE: Rightmost column shows the *p* value returned by a Fisher's exact test; small values indicate rejection of the null hypothesis of no association between sex and presence of disease.

12.4. Medication usage variables

Summary information for the two dichotomous indicators of medication use is given in Table 12.4.1.

TABLE 12.4.1: Observed prevalence values for use of two specific types of medication. Cell content is number of positives (individuals reporting use), followed by % of positives, followed by total number of cases.

medication type	**MALES** ***n* positive / % prevalence (total *n*)**	**FEMALES** ***n* positive / % prevalence (total *n*)**	**ALL** ***n* positive / % prevalence (total *n*)**	***p***
anti-hypertensive	41 / 29.5 (139)	44 / 28.6 (154)	85 / 29.0 (293)	0.90
bronchodilator	13 / 9.4 (139)	17 / 11.0 (154)	30 / 10.2 (293)	0.70

NOTE: Rightmost column shows the *p* value returned by a Fisher's exact test; small values indicate rejection of the null hypothesis of no association between sex and medication use.

CHAPTER 13: RESULTS (III) - ASSOCIATIONS LINKING SOCIAL POSITION, RESIDENTIAL CONDITIONS AND HEALTH

13.1 Introduction

This chapter reports on the main associations which were investigated with the aim of assessing the three realisations of the general conceptual model postulated in the study (identified in the *Introduction* as Models A, B and C). These associations were summarised in Section 9.1. Findings relating to these associations are now described in an order which seeks to impose a logical structure on the large number of individual analyses involved. Three basic principles are observed. First, associations which involve health are treated before the remaining relationships (i.e. those which link the two other factors of interest, namely SEP and residential conditions). This reflects the fact that the primary focus of the study is on investigating certain hypothesised determinants of health. Second, associations between accumulated exposures and health are reported before those which link time-dependent (clustered) experiences and health. Broadly speaking, the rationale behind this ordering is to separate, so far as is possible, consideration of accumulation-of-risk effects from critical period influences. Third, relationships which involve the physiological health outcomes (blood pressure and lung function) are treated first, followed by associations which relate to the experience of specific disease types (heart disease, lung disease etc.). Associations involving use of medications (bronchodilator / antihypertensive) are then considered. This reflects a basic chronological and causal ordering: physiological status (where judged abnormal) leads to the clinical diagnosis of disease, which in turn generates medication use.

Because the number of individual associations to be reported is large, presenting the findings in an intelligible and informative way is challenging. To assist assimilation of these results, it is helpful to identify (a) the conceptual model(s) to which each result is relevant (i.e. Model A, B and / or C) and (b) the analytical regime used to derive the result (see *Methods*, Sections 9.3 to 9.8). Doing so helps to maintain awareness of both *why* and *how* each association was investigated. Table 13.1.1 *(next page)* provides this information as part of a general outline of the structure of the present chapter. In order to further assist appreciation of the findings, the chapter concludes with a brief summary of the significant effects identified (Section 13.18).

TABLE 13.1.1: Chapter structure. Column headed 'model(s)' identifies the conceptual model(s) to which each set of associations relates. Column headed 'analytical regime' identifies the analytical regime used to investigate each set of associations, followed (in brackets) by the section in the *Methods* chapters which describes that regime.

section	**associations reported**	**model(s)**	**analytical regime (*Methods* section)**
13.2	Accumulated disadvantage *with* physiological variables	B, C	1 (9.3)
13.3	Accumulated exposure to residential hazards *with* physiological variables	B	1 (9.3)
13.4	Accumulated disadvantage *with* clinical variables	B, C	2 (9.4)
13.5	Accumulated exposure to residential hazards *with* clinical variables	B	2 (9.4)
13.6	Accumulated disadvantage *with* medication usage variables	B, C	2 (9.4)
13.7	Accumulated exposure to residential hazards *with* medication usage variables	B	2 (9.4)
13.8	Accumulated exposures with secondary health outcomes	N / A	2 (9.4)
13.9	Time-dependent social location *with* physiological variables	A	3 (9.5)
13.10	Time-dependent exposure to residential hazards *with* physiological variables	A, C	3 (9.5)
13.11	Time-dependent social location *with* clinical variables	A	4 (9.6)
13.12	Time-dependent exposure to residential hazards *with* clinical variables	A, C	4 (9.6)
13.13	Time-dependent social location *with* medication usage variables	A	4 (9.6)
13.14	Time-dependent exposure to residential hazards *with* medication usage variables	A, C	4 (9.6)
13.15	Time-dependent influences *with* secondary health outcomes	N / A	4 (9.6)
13.16	Accumulated disadvantage *with* accumulated exposure to residential hazards	B, C	5 (9.7)
13.17	Time-dependent social location with time-dependent exposure to residential hazards	A	6 (9.8)

13.2 Associations (i): accumulated disadvantage and the physiological variables

Table 13.2.1 *(next page)* shows, for each of the three physiological measures, the rank correlation of that measure with accumulated disadvantage. The table indicates that while no meaningful association was observed between accumulated disadvantage and either measure of blood pressure, there was evidence of a modest inverse relationship between accumulated disadvantage and lung function.

TABLE 13.2.1: Rank correlations of accumulated disadvantage with three physiological measures. Cell content is correlation coefficient (with 95% confidence interval) followed by number of cases from which coefficient is derived.

rank correlation of accumulated disadvantage with…	**r_s (95% CI) / *n***
systolic blood pressure (mm Hg)	0.01 (-0.11 to 0.12) / 282
diastolic blood pressure (mm Hg)	-0.01 (-0.13 to 0.11) / 282
standardised FEV_1 (%)	-0.17 (-0.28 to -0.05) 282

Results from the regression models fitted to assess these associations in greater details are presented in Table 13.2.2. This shows the parameter estimates for the effect of one year's experience of social disadvantage on each health measure. The table confirms the presence of a significant negative association between accumulated disadvantage and standardised FEV_1.

TABLE 13.2.2: Estimated effect of one year's experience of social disadvantage on three physiological measures. Estimates are adjusted for the effects of sex and smoking status.

health measure	**estimate (95% CI)**
systolic blood pressure (mm Hg)	-0.01 (-0.18 to 0.15)
diastolic blood pressure (mm Hg)	-0.04 (-0.12 to 0.04)
standardised FEV_1 (%)	-0.11 (-0.19 to -0.03)

13.3 Associations (ii): accumulated exposure to residential hazards and the physiological variables

The rank correlations between the three measures of cumulative exposure to residential risks and the physiological health measures are given in Table 13.3.1. The confidence intervals around all coefficients include the value zero, and are thus consistent with no significant association being present.

TABLE 13.3.1: Rank correlations of three measures of accumulated exposure to adverse residential conditions with three physiological measures. Cell content is correlation coefficient (with 95% confidence interval) followed by number of cases from which coefficient is derived.

hazard measure	**rank correlation: systolic BP (mm Hg) / *n***	**rank correlation: diastolic BP (mm Hg) / *n***	**rank correlation: standardised FEV_1 (%) / *n***
damp exposure (yrs)	-0.04 (-0.17 to 0.09) 219	-0.10 (-0.23 to 0.04) 219	-0.03 (-0.16 to 0.11) 219
air pollution exposure (yrs)	0.05 (-0.08 to 0.19) 214	0.05 (-0.08 to 0.18) 214	0.00 (-0.13 to 0.13) 214
total hazard load (yrs)	-0.01 (-0.14 to 0.13) 209	-0.03 (-0.17 to 0.10) 209	-0.05 (-0.19 to 0.08) 209

Table 13.3.2 *(next page)* presents results from the associated regression models, showing the estimated effect of an additional year's exposure on each outcome. Scrutiny of the

confidence intervals around the parameter estimates indicates that none of the effects is statistically significant at the conventional 5% level.

TABLE 13.3.2: Estimated effect of one year's exposure to three residential hazards on three physiological measures. Estimates are adjusted for the effects of sex and smoking status.

hazard measure	**effect on: systolic BP (mm Hg)**	**effect on: diastolic BP (mm Hg)**	**effect on: standardised FEV_1 (%)**
damp exposure (yrs)	-0.12 (-0.40 to 0.21)	-0.03 (-0.20 to 0.21)	-0.08 (-0.23 to 0.04)
air pollution exposure (yrs)	0.04 (-0.21 to 0.35)	0.02 (-0.11 to 0.19)	-0.03 (-0.15 to 0.09)
total hazard load (yrs)	0.00 (-0.19 to 0.24)	0.02 (-0.09 to 0.17)	-0.03 (-0.12 to 0.05)

13.4 Associations (iii): accumulated disadvantage and the clinical variables

Table 13.4.1 shows odds ratio estimates for the effect of an additional year spent in disadvantage on each of the four indicators representing the presence of specific disease types. The table indicates that accumulated disadvantage is significantly associated with elevated odds of stroke.

TABLE 13.4.1: Odds ratio estimates for the effect of accumulated disadvantage (one additional year of exposure) on presence of four specific disease types. Estimates are adjusted for the effects of sex and smoking status.

disease type	**odds ratio (95% CI)**
heart disease	1.00 (0.98 to 1.01)
lung disease	1.02 (0.99 to 1.04)
stroke	1.06 (1.01 to 1.12)
high blood pressure	1.01 (0.99 to 1.02)

13.5 Associations (iv): accumulated exposure to residential hazards and the clinical variables

Results from logistic regression models predicting the effect of exposure to residential hazards on the clinical variables are presented in Table 13.5.1 *(next page)*. With one exception (the association between dampness exposure and lung disease), the confidence intervals for all effects include the value one which indicates equal odds for both outcome states.

TABLE 13.5.1 Odds ratio estimates for the effect of accumulated exposure to three residential hazards (one additional year of exposure) on the presence of four specific disease types. Estimates are adjusted for the effects of sex and smoking status.

disease type	**effect of damp exposure: odds ratio (95% CI)**	**effect of air pollution exposure: odds ratio (95% CI)**	**effect of total hazard load: odds ratio (95% CI)**
heart disease	1.02 (0.99 to 1.05)	1.01 (0.98 to 1.03)	1.01 (0.99 to 1.03)
lung disease	1.04 (1.01 to 1.08)	1.02 (0.98 to 1.05)	1.02 (0.99 to 1.04)
stroke	1.03 (0.98 to 1.08)	1.02 (0.98 to 1.06)	1.03 (0.99 to 1.06)
high blood pressure	1.01 (0.98 to 1.04)	0.99 (0.97 to 1.02)	1.00 (0.98 to 1.02)

13.6 Associations (v): accumulated disadvantage and the medication usage variables

Table 13.6.1 gives estimated odds ratios for the effect of accumulated disadvantage on the binary indicators representing the use of specific types of medication. The table indicates that accumulated disadvantage displays a marginally significant association with use of bronchodilator medication; the significance of the parameter is $p = 0.06$[38].

TABLE 13.6.1 Odds ratio estimates for the effect of accumulated disadvantage (one additional year of exposure) on the use of two specific types of medication. Estimates are adjusted for the effects of sex and smoking status.

medication type	**odds ratio (95% CI)**
anti-hypertensive	1.00 (0.99 to 1.02)
bronchodilator	1.03 (1.00 to 1.05)

13.7 Associations (vi): accumulated exposure to residential hazards and the medication usage variables

The estimated effect of cumulative exposure to residential hazards on medication use is reported in Table 13.7.1 *(next page)*. The confidence limits around all estimates include the value one, and are thus consistent with no significant association being present.

[38] Throughout this thesis, the result of a hypothesis test is considered 'marginal' if the *p* value slightly exceeds the conventional significance limit of 0.05 (or 5%). Similarly, a confidence interval around an estimate is deemed marginal if it narrowly includes the value consistent with no effect or association. While imprecise, this approach avoids treating the traditional 5% significance level as a rigid threshold, above which a result is instantly deemed to indicate a complete absence of effect. Such an interpretation is clearly incorrect: "It is ridiculous to interpret the results of a study differently according to whether the P value obtained was, say, 0.055 or 0.045." (Altman 1991, p. 168).

TABLE 13.7.1 Odds ratio estimates for the effect of accumulated exposure to residential hazards (one additional year of exposure) on the use of two specific types of medication. Estimates are adjusted for the effects of sex and smoking status.

medication type	**effect of damp exposure: odds ratio (95% CI)**	**effect of air pollution exposure: odds ratio (95% CI)**	**effect of total hazard load: odds ratio (95% CI)**
anti-hypertensive	1.02 (0.99 to 1.04)	0.98 (0.96 to 1.01)	1.00 (0.98 to 1.02)
bronchodilator	1.03 (0.99 to 1.07)	1.02 (0.98 to 1.05)	1.01 (0.99 to 1.04)

13.8 Associations (vii): accumulated exposures and the secondary health outcomes

The associations between cumulative disadvantage and the secondary health outcomes were nonsignificant, all *p* values being 0.20 or greater. With one exception, associations between the three measures of residential risk and the secondary outcomes were also nonsignificant; *p* values for these were 0.11 or greater. The single exception was a marginally significant relationship between total hazard load and reported thyroid disease. The estimated odds ratio for this effect was 0.95 (95% CI: 0.90 to 1.00; $p = 0.06$).

13.9 Associations (viii): time-dependent social location and the physiological variables

Table 13.9.1 shows the results of Kruskal-Wallis tests applied to the three physiological health outcomes, with the data classified by social position group (i.e. the clustered scheme of Appendix 2). All *p* values exceed the conventional 5% level for statistical significance, indicating non-rejection of the null hypothesis[39] .

TABLE 13.9.1: Results from Kruskal-Wallis tests applied to three physiological measures, with subjects classified by social location cluster.

health measure	***p***
systolic blood pressure (mm Hg)	0.99
diastolic blood pressure (mm Hg)	0.84
standardised FEV1 (%)	0.13

13.10 Associations (ix): time-dependent exposure to residential hazards and the physiological variables

Table 13.10.1 *(next page)* presents the *p* values returned by Kruskal-Wallis tests applied to the physiological measures, respondents being classified by the grouped dampness scheme of Appendix 3.

[39] See Section 9.5 for comments by Kruskal and Wallis on the interpretation of the test.

TABLE 13.10.1: Results from Kruskal-Wallis tests applied to three physiological measures, with subjects classified by dampness cluster.

health measure	***p***
systolic blood pressure (mm Hg)	0.05
diastolic blood pressure (mm Hg)	0.13
standardised FEV1 (%)	0.75

Because the test for systolic blood pressure indicated marginal rejection of the null hypothesis, this association was investigated further via a multiple regression model in which systolic BP was predicted by dampness group membership, together with sex and smoking status. The dampness classes were represented in the models by a family of five dummy variables indicating respectively membership of Clusters 2 to 6 (see Appendix 3). Thus, the parameter estimates for these indicators represent the change in the outcome associated with membership of each dampness group relative to the zero-exposure Cluster 1. The estimates associated with these indicators are shown in Table 13.10.2, which indicates that systolic blood pressure exhibited some variation across dampness groups.

TABLE 13.10.2: Estimated effect of dampness group membership on systolic blood pressure; estimates expressed relative to dampness Cluster 1 (zero exposure). Cell content is estimate, followed by 95% confidence interval. Estimates are adjusted for the effects of sex and smoking status.

dampness cluster	**effect on systolic BP (mm Hg): estimate (95% CI)**
2	0.3 (-7.3 to 9.3)
3	-15.8 (-25.5 to -3.4)
4	4.2 (-5.7 to 19.5)
5	-2.6 (-16.3 to 26.2)
6	17.9 (4.8 to 46.5)

The results of Kruskal-Wallis tests applied to the physiological measures, with the data classified by air pollution group (see Appendix 4), are given in Table 13.10.3. These results indicate that, for all three health measures, there is insufficient evidence to reject the null hypothesis.

TABLE 13.10.3: Results from Kruskal-Wallis tests applied to three physiological measures, with subjects classified by air pollution cluster.

health measure	***p***
systolic blood pressure (mm Hg)	0.20
diastolic blood pressure (mm Hg)	0.10
standardised FEV1 (%)	0.12

13.11 Associations (x): time-dependent social location and the clinical variables

Table 13.11.1 presents results from Freeman-Halton tests of association[40] between the four indicators representing the presence of specific disease types and the clustered representation of social location. Results for heart disease, lung disease and high blood pressure are nonsignificant, indicating an absence of meaningful association; however, a significant relationship is suggested for stroke.

TABLE 13.11.1: Results from Freeman-Halton tests of association between social location cluster scheme and the presence of four specific disease types.

disease type	***p* (99% CI)**
heart disease	0.99 (0.98 to 0.99)
lung disease	0.35 (0.34 to 0.37)
stroke	0.01 (0.01 to 0.02)
high blood pressure	0.53 (0.52 to 0.54)

13.12 Associations (xi): time-dependent exposure to residential hazards and the clinical variables

Results from Freeman-Halton tests of association between the four disease indicators and the clustered representations of exposure to dampness and air pollution are given in Table 13.12.1 *(next page)*. All results are nonsignificant at the conventional 5% level, although a marginally significant relationship between dampness exposure and the experience of lung disease is suggested.

[40] As explained in *Methods* Section 9.6, these results were obtained via Monte Carlo simulation because the time required to compute the results of true F-H tests would be prohibitive. This applies to all instances of F-H tests reported in this thesis. In Table 13.11.1, and in other tables which report the results of F-H tests, the confidence intervals associated with each *p* value quantify the precision with which the Monte Carlo estimate represents the (unknown) exact *p* value.

TABLE 13.12.1: Results from Freeman-Halton tests of association between cluster schemes for dampness and air pollution exposure, and the presence of four specific disease types.

disease type	**DAMPNESS** ***p*** **(99% CI)**	**AIR POLLUTION** ***p*** **(99% CI)**
heart disease	0.40 (0.39 to 0.41)	0.76 (0.74 to 0.77)
lung disease	0.06 (0.05 to 0.06)	0.33 (0.32 to 0.35)
stroke	0.17 (0.16 to 0.18)	0.17 (0.16 to 0.18)
high blood pressure	1.00 (1.00 to 1.00)	0.83 (0.82 to 0.84)

13.13 Associations (xii): time-dependent social location and the medication usage variables

Table 13.13.1 gives the results of Freeman-Halton tests of association between the clustered representation of social position and the indicators representing medication use. For both types of medication, the results indicate that use is not related to respondents' time-related social position.

TABLE 13.13.1: Results from Freeman-Halton tests of association between social location cluster scheme and the use of two specific types of medication.

medication type	***p*** **(99% CI)**
anti-hypertensive	0.75 (0.74 to 0.76)
bronchodilator	0.22 (0.20 to 0.23)

13.14 Associations (xiii): time-dependent exposure to residential hazards and the medication usage variables

The values returned by Freeman-Halton tests of association between the clustered representations of exposure to residential hazards and medication use are shown in Table 13.14.1. The results of these tests indicate that none of the associations is significant at the conventional 5% level.

TABLE 13.14.1: Results from Freeman-Halton tests of association between cluster schemes for dampness and air pollution exposure, and the use of two specific types of medication.

medication type	**DAMPNESS** ***p*** **(99% CI)**	**AIR POLLUTION** ***p*** **(99% CI)**
anti-hypertensive	0.70 (0.69 to 0.72)	0.73 (0.71 to 0.74)
bronchodilator	0.30 (0.29 to 0.31)	0.34 (0.33 to 0.36)

13.15 Associations (xiv): time-dependent influences and the secondary health outcomes

Associations between the three cluster schemes and each of the ten secondary health outcomes were also assessed via Freeman-Halton tests. Of the 30 tests performed, the only result judged significant at the conventional 5% level was that representing the association between time-dependent social position and the presence of long-standing illness ($p = 0.01$).

13.16 Associations (xv): accumulated disadvantage and accumulated exposure to residential hazards

Table 13.16.1 shows the estimated effect of one year's experience of social disadvantage on each of the three measures of accumulated exposure to residential hazards. Estimates were derived from multiple regression models of the form shown in Section 9.7. The table indicates that cumulative disadvantage exhibits significant positive associations with exposure to air pollution and with total hazard load.

TABLE 13.16.1: Estimated effect of one year's experience of social disadvantage on three measures of exposure to adverse residential conditions. Estimates are adjusted for the effect of gender.

hazard measure	**estimate (95% CI)**
damp exposure (yrs)	0.06 (-0.02 to 0.15)
air pollution exposure (yrs)	0.15 (0.06 to 0.26)
total hazard load (yrs)	0.23 (0.10 to 0.36)

13.17 Associations (xvi): time-dependent social location and time-dependent exposure to residential hazards

Freeman-Halton tests of association between the clustered representation of social position and the grouped representations of exposure to residential hazards returned $p = 0.78$ (95% CI: $p = 0.77$ to $p = 0.79$) for dampness, and $p = 0.10$ (95% CI: $p = 0.09$ to $p = 0.11$) for air pollution. There is thus insufficient evidence from these data to conclude that time-dependent social position is related to the experience over time of residential dampness or exposure to air pollution.

***

In order to assist in overall interpretation of the large number of individual associations reported above, the chapter now concludes with a summary of the findings.

13.18 Summary of observed associations

13.18.1 Accumulated disadvantage and health

Cumulative disadvantage was found to exhibit a significant inverse relationship with lung function (Section 13.2) and a positive relationship with the reported experience of stroke (Section 13.4). A marginally significant ($p = 0.06$) association of disadvantage with use of bronchodilator medication was indicated (Section 13.6).

13.18.2 Accumulated exposure to residential hazards and health

No evidence was found of significant associations between the measures of cumulative residential hazard exposure and the physiological variables (Section 13.3). Accumulated exposure to residential dampness was associated with elevated odds of reporting lung disease (Section 13.5). No association between cumulative residential exposures and medication use was found (Section 13.7). A marginally significant ($p = 0.06$) relationship of total hazard load with reduced odds of reporting thyroid disease (a secondary health outcome) was observed (Section 13.8).

13.18.3 Time-dependent social location and health

The grouped representation of time-ordered social position was not found to exhibit a significant relationship with any of the three physiological variables (Section 13.9). Clustered social position was significantly associated with the reported experiences of stroke (Section 13.11) and of long-standing illness (Section 13.15), the latter being a secondary health outcome. No relationship of time-dependent social position with medication use was indicated (Section 13.13).

13.18.4 Time-dependent exposure to residential hazards and health

The clustered experience of exposure to dampness demonstrated a significant relationship with systolic blood pressure (Section 13.10) and a marginally significant ($p = 0.06$) association with reported lung disease (Section 13.12). Associations between time-

dependent exposure to residential hazards and medication use were not significant (Section 13.14).

13.18.5 Accumulated disadvantage and accumulated exposure to residential hazards

Accumulated disadvantage exhibited significant positive associations with cumulative exposure to air pollution and with total hazard load (Section 13.16).

13.18.6 Time-dependent social location and time-dependent exposure to residential hazards

The clustered representation of social position was not significantly associated with either of the grouped representations of residential hazard exposure (Section 13.17).

***

The presentation of results continues by examining the properties of subjects' detailed sequences of experience over time.

CHAPTER 14: RESULTS (IV) - PROPERTIES OF THE DETAILED SEQUENCES OF SOCIAL POSITION AND RESIDENTIAL CONDITIONS

14.1 Introduction and chapter structure

This chapter presents results derived from exploratory investigation of the sequences or trajectories of year-on-year experience, using mainly the methods outlined in Chapter 10. The sets of trajectories themselves are too large for convenient inclusion within the text of this chapter, and are therefore shown in appendices as follows:-

socioeconomic position	-	Appendix 5
dampness exposure	-	Appendix 6
air pollution exposure	-	Appendix 7

The correspondence between the methods described in Chapter 10, and the material presented in the present chapter, is shown in Table 14.1.1.

TABLE 14.1.1: Chapter structure

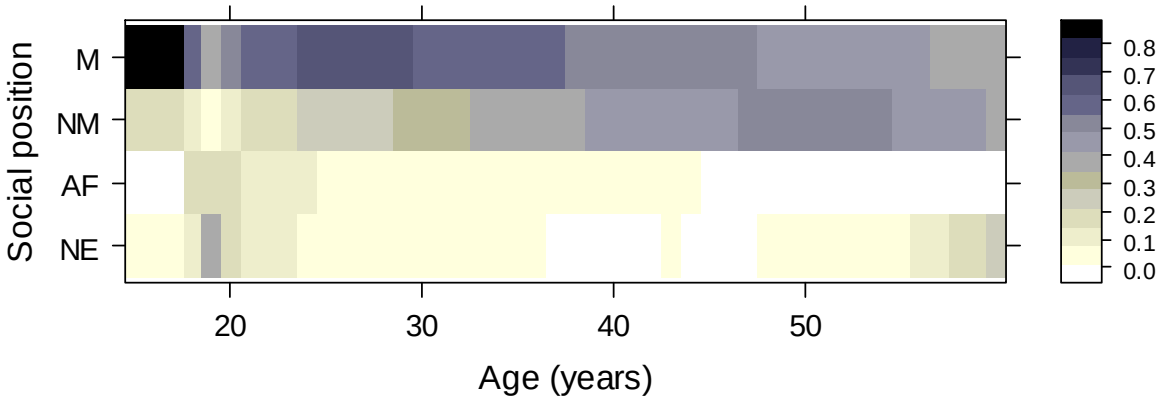
section	**type of result**	**related section in *Methods* (Chapter 10)**
14.2	graphical position weight matrices (representing trajectories of socioeconomic position)	10.1
14.3	'spike' charts (representing trajectories of residential hazard exposure)	10.2
14.4	*k*-gram summaries of transition events (representing trajectories of social position)	10.3
14.5	graphical position weight matrices (representing trajectories of combined exposure)	10.4

A final section of this chapter (14.6) introduces three numeric measures (not covered in Chapter 10) which were developed to summarise certain general properties of the sets of social and residential hazard trajectories. These measures are introduced at this stage rather than in the *Methods* section because they are most easily illustrated by referring to actual results from the study (which by convention are deemed not be available when the methods are expounded).

14.2 Socioeconomic position over time - graphical position weight matrices

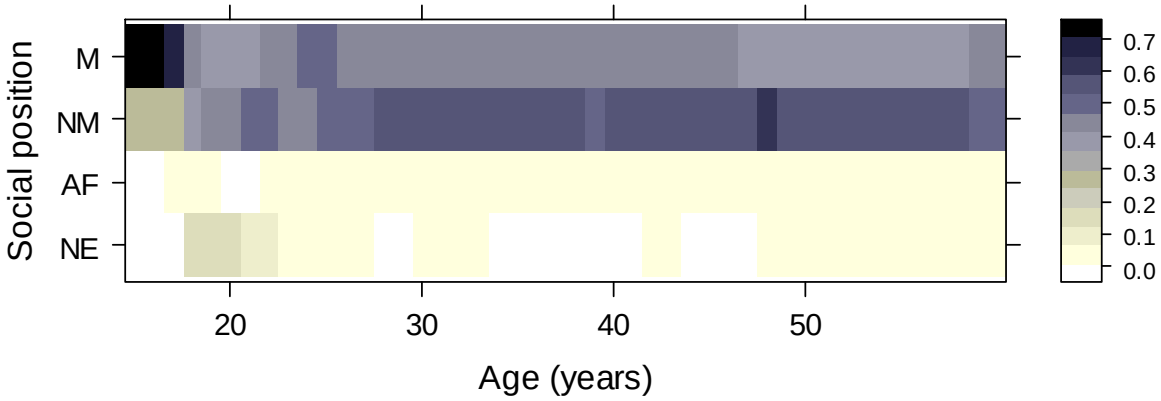
Graphical position weight matrices representing subjects' trajectories of SEP over time are shown in Figure 14.2.1a (males) and 14.2.1b (females).

FIGURE 14.2.1a: Position weight matrix representing trajectories of SEP over the age range 15-60 years for 139 male subjects. Explanation of symbols on vertical axis appears beneath figure. Shades indicate the proportion of the total number of subjects in each social state at each age point.



NOTE: Symbols used for social position (vertical axis) are: 'M' = manual, 'NM' = non-manual, 'AF' = Armed Forces, 'NE' = non-employed.

FIGURE 14.2.1b: Position weight matrix representing trajectories of SEP over the age range 15-60 years for 155 female subjects. Explanation of symbols on vertical axis appears beneath figure. Shades indicate the proportion of the total number of subjects in each social state at each age point.



NOTE: Symbols used for social position (vertical axis) are: 'M' = manual, 'NM' = non-manual, 'AF' = Armed Forces, 'NE' = non-employed.

14.3 Experience of residential hazards over time - exposure charts

Figure 14.3.1 *(next page)* shows the proportion of individuals exposed to residential dampness at each yearly age point in the range covered by the study. A corresponding display for air pollution exposure is given as Figure 14.3.2 *(next page)*.

FIGURE 14.3.1: Proportion (percentage) of individuals recording exposure to dampness at each year in the age range 15-60 years (subjects classified by sex; *n* = 97 [male], *n* = 123 [female]).

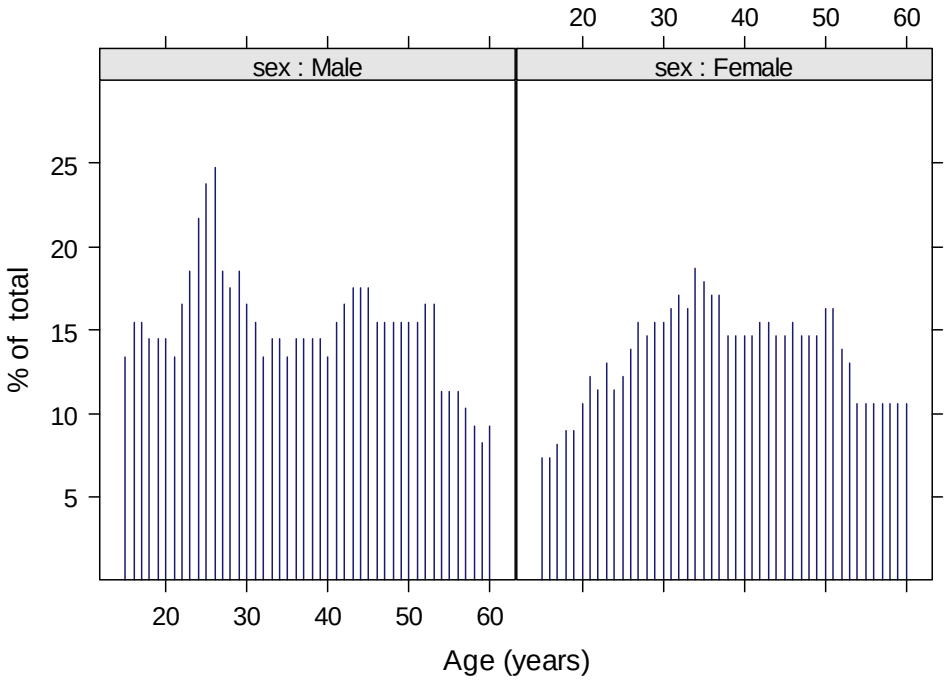
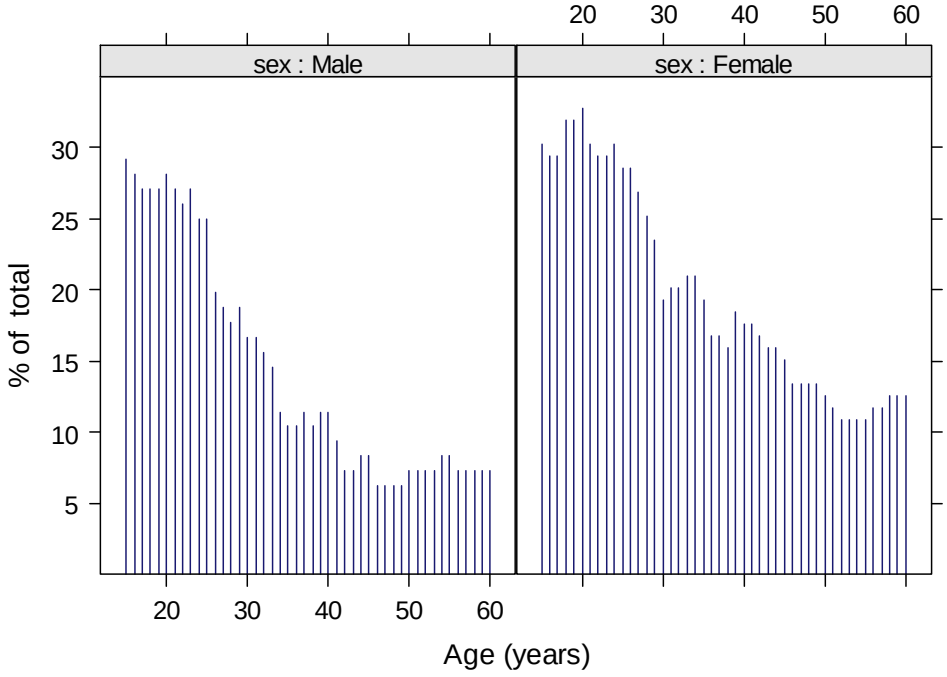


FIGURE 14.3.2: Proportion (percentage) of individuals recording exposure to air pollution at each year in the age range 15-60 years (subjects classified by sex; *n* = 96 [male], *n* = 119 [female]).



14.4 Transition patterns in the sequences of socioeconomic position

14.4.1 Prevalence of individual transition types

Table 14.4.1 shows the respective frequencies with which the 12 possible types of social transition (e.g. *from* manual at age *Y* years *to* non-manual at age *Y*+1) were observed in the data.

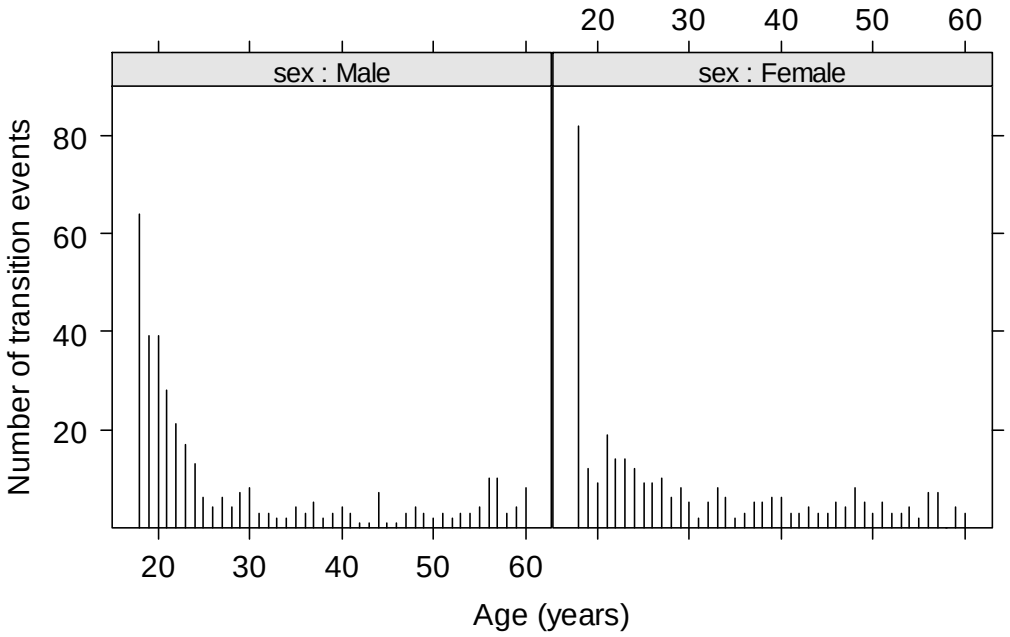
TABLE 14.4.1: Frequency of occurrence of 12 types of change in socioeconomic position between age *Y* years and age *Y*+1. Cell content is number of instances of transition type *X* (first element), followed by percentage of all transitions (second element; column percentages sum to 100). Values are shown separately for (from left) males, females and all respondents.

type of status change	**MALE (*n* = 139) no. (% of total)**	**FEMALE (*n* = 155) no. (% of total)**	**ALL SUBJECTS (n = 294) no. (% of total)**
manual ➔ non-manual	66 (18.2)	135 (40.1)	201 (28.7)
manual ➔ non-employed	79 (21.8)	18 (5.3)	97 (13.9)
manual ➔ Armed Forces	26 (7.2)	2 (0.6)	28 (4.0)
non-manual ➔ manual	37 (10.2)	100 (29.7)	137 (19.6)
non-manual ➔ non-employed	32 (8.8)	30 (8.9)	62 (8.9)
non-manual ➔ Armed Forces	5 (1.4)	5 (1.5)	10 (1.4)
non-employed ➔ manual	50 (13.8)	9 (2.7)	59 (8.4)
non-employed ➔ non-manual	27 (7.4)	33 (9.8)	60 (8.6)
non-employed ➔ Armed Forces	5 (1.4)	0 (0.0)	5 (0.7)
Armed Forces ➔ manual	21 (5.8)	2 (0.6)	23 (3.3)
Armed Forces ➔ non-manual	14 (3.9)	2 (0.6)	16 (2.3)
Armed Forces ➔ non-employed	1 (0.3)	1 (0.3)	2 (0.3)
ALL TYPES	363 (100.0)	337 (100.0)	700 (100.0)

14.4.2 Transition activity at specific ages

Figure 14.4.1 *(next page)* shows, in graphical form, the absolute numbers of social state transitions (all types combined) observed at each yearly age point.

FIGURE 14.4.1: Numbers of social transition events (all types combined) observed at yearly age points in the range 18-60 years [see Note below figure] (subjects classified by sex; *n* = 139 [male], *n* = 155 [female]).

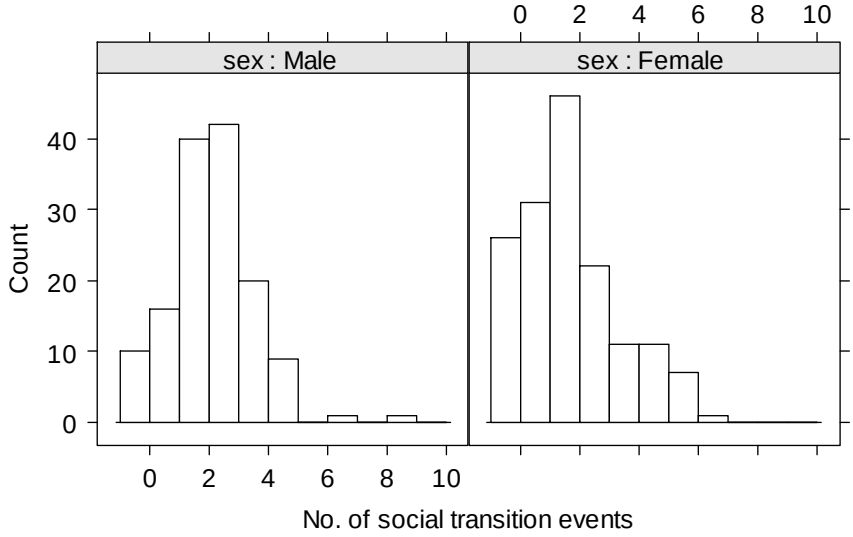


NOTE: Because parental occupational class is assumed to apply continuously until age 17, transitions are possible only from age 18 onwards. Transitions are allocated to the later of the two years involved e.g. a transition from manual status at age 21 to non-manual at age 22 is assigned to age 22.

14.4.3 Distribution of the numbers of transitions observed for subjects

Figure 14.4.2 shows the overall distribution of the number of observed transition events.

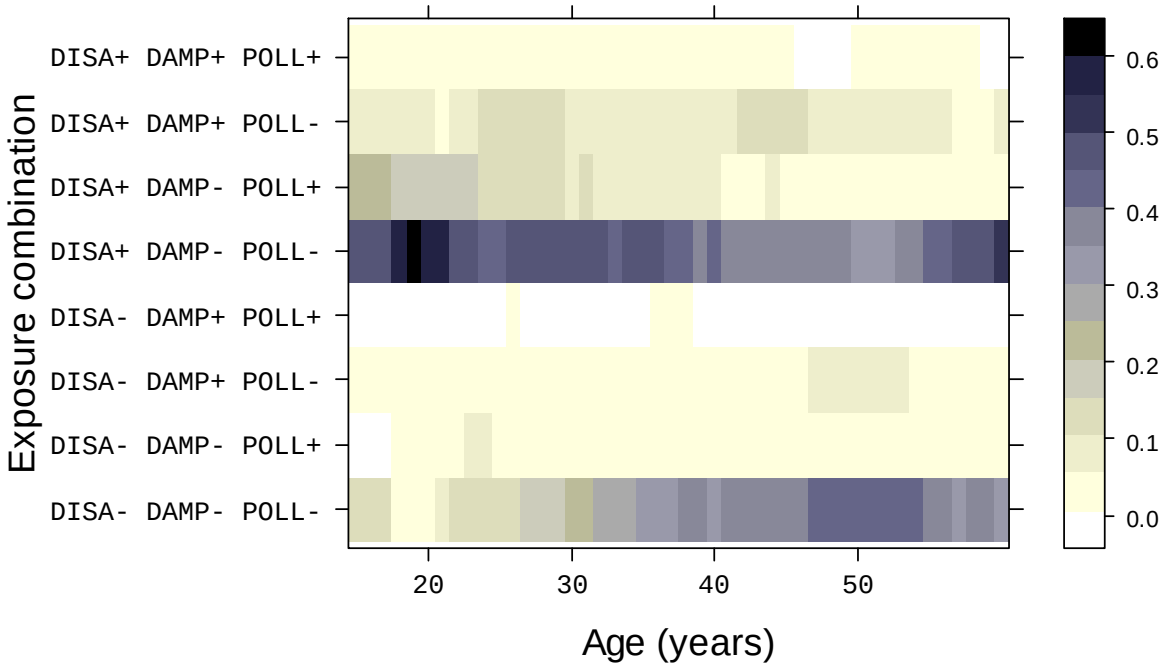
FIGURE 14.4.2: Distribution of number of social transition events observed over the age range 18-60 years (subjects classified by sex; *n* = 139 [male], *n* = 155 [female]).



14.5 Exposure to combined hazards over time

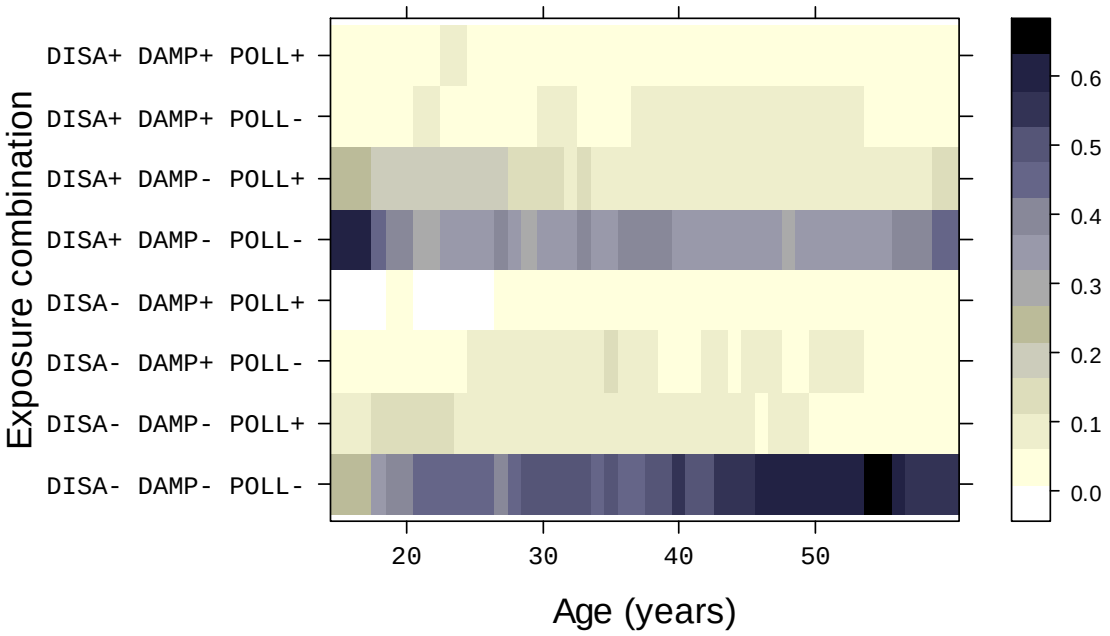
A method for representing subjects' exposure to combinations of hazards over time was described in Section 10.4. Graphical position weight matrices generated via this method are presented in Figure 14.5.1a (males) and 14.5.1b (females *[next page]*).

FIGURE 14.5.1a: Position weight matrix showing combined exposure by age to three hazard types over the age range 15-60 years (male subjects; *n* = 94). Explanation of symbols on vertical axis appears beneath figure. Shades indicate the proportion of the total number of subjects experiencing each combination at each age point.



NOTE: Hazard combinations are identified on the vertical axis by the abbreviations DISA(dvantage), DAMP(ness) and POLL(ution). A plus symbol ('+') after each hazard indicates exposure to that hazard, a minus symbol ('-') denotes freedom from exposure to that hazard.

FIGURE 14.5.1b: Position weight matrix showing combined exposure by age to three hazard types over the age range 15-60 years (female subjects; *n* = 116). Explanation of symbols on vertical axis appears beneath figure. Shades indicate the proportion of the total number of subjects experiencing each combination at each age point.



NOTE: Hazard combinations are identified on the vertical axis by the abbreviations DISA(dvantage), DAMP(ness) and POLL(ution). A plus symbol ('+') after each hazard indicates exposure to that hazard, a minus symbol ('-') denotes freedom from exposure to that hazard.

14.6 Characterising sequence sets - commonality and complexity

The presentation of results relating to the detailed sequences shown in Appendices 5 to 7 concludes by introducing two concepts which are helpful in summarising certain characteristics of these sets of trajectories, and will feature prominently in the *Discussion*. The justification for developing these ideas at this point (rather than in the *Methods* section) was given in Section 14.1. The first of these concepts, henceforth termed *commonality*[41], is defined as whether an observed trajectory is common to multiple respondents, or is unique to a single individual. The idea may be illustrated with reference to Appendix 5. This shows that the trajectory indicating persistent manual status (identified as sequence ID 1) was the most frequently observed (being shared by 30 subjects), while a further 22 sequences (ID 2 to ID 23) were also recorded for multiple individuals. Conversely, all sequences from ID 24 onwards were unique: that is, represented the experience of only a single person. Commonality as defined here is that binary property of an individual trajectory which

[41] "The state or quality of being in common with, or shared by, others... ...community of function, structure, or purpose; also, a shared feature" (Oxford English Dictionary, 2009)

indicates whether it was nonunique or unique. The concept may trivially be extended to represent the number of subjects who share a common sequence (e.g. 30 individuals in the case of sequence ID 1).

A second property of an individual sequence which is helpful in discussion will be referred to as *complexity*[42]. While this term is widely used in scientific contexts, it is employed here in a specific sense which is now defined, the concept again being developed with reference to Appendix 5. Sequence ID 1 (representing persistent manual status) is clearly 'less complex' (in the everyday sense of the word) than sequence ID 2, which depicts an initial period of three 'manual' years followed by a uniform experience of the non-manual condition over the remainder of the period examined. In turn, ID 2 is less complex than ID 25, which includes periods spent in three different states: manual (three times), non-employed (three times) and engaged in Armed Forces service (once). The concept of complexity is one which persistently recurs in numerous areas of science (albeit with very different meanings), and has been defined in many ways. To illustrate, around 40 proposed measures of complexity have been listed by Lloyd (2001). Some of the existing measures of complexity used in other disciplines (such as statistical complexity [Crutchfield & Young, 1989] or Kolmogorov complexity [Li & Vitanyi, 2008]) might feasibly be adapted to characterise the complexity of sequences such as those of Appendix 5. In a context more closely related to that of the present study, Elzinga (2009) has proposed measures of the complexity of categorical time series which are based on the number of distinct subsequences observed[43]. However, for the purposes of the present discussion a simpler approach is adopted, based on enumerating the number of transitions or changes of state which are present in each sequence. Remaining with the examples discussed thus far, sequence ID 1 contains no transitions; ID 2 includes a single transition (*from* manual *to* non-manual); and no fewer than six state changes are evident in ID 25. In the present discussion, the complexity of a single sequence is defined as the number of state changes it contains. Complexity in this sense may be thought of as a measure of flux; that is, the extent to which the sequential experience of some factor (here, social position) is characterised by frequent changes in state or (alternatively) is stable over time.

[42] "Involved nature or structure, intricacy" (Oxford English Dictionary, 2009)

[43] Elzinga's paper concentrates particularly on complexity in the lifecourse in the field of sociology, and is thus directly related to one of the themes of this thesis. However, the paper was unpublished (but in press) as at August 2009 when the first draft of the thesis was completed, and ideas in the thesis relating to sequence complexity were developed independently, before the author obtained sight of Elzinga's paper.

The two concepts introduced above – commonality and complexity – are fundamentally different. By indicating the uniqueness status of an observed trajectory, commonality says something about an individual subject's relationship to other subjects in the dataset i.e. whether her / his pattern of experience is or is not shared by others. Conversely, complexity expresses one aspect of the internal structure of a sequence, taking no account of whether the experience represented by that sequence is shared or not. Both concepts may usefully be applied to characterise *sets* of sequences (such as those created for this study), via simple measures which are now described.

The commonality characteristics of a sequence set such as that of Appendix 5 may be expressed quantitatively in a number of ways, two of which will feature in the present discussion. The first of these (referred to hereafter as the *mean sequence frequency* [MSF]) is simply the average number of individuals represented by a single sequence in the set. Appendix 5 shows that 216 distinct trajectories were identified among the 294 individuals in the dataset, so the MSF is given by

$$\frac{294}{216} = 1.36$$

This quantity is subject to a lower bound of 1 (indicating that all sequences in the set are unique), and a higher bound equal to the number of cases in the dataset. The latter value would indicate that all subjects shared a common sequence. Generally, higher values for the MSF indicate greater commonality (or lower levels of uniqueness) in the set.

A second measure of commonality over a set of sequences is provided by the proportion (or percentage) of the total number of individual subjects who are represented by unique trajectories; that is, whose sequence of experience is not shared with any other subject. From Appendix 5, 193 of 294 respondents recorded unique sequences of social position, so the value of this second measure (termed the *sequence uniqueness index* [SUI]) for these data is

$$\frac{193}{294} = 0.66 \text{ (or 66\%)}$$

Values of this quantity are bounded by zero (indicating that no subjects have a unique trajectory) and by 1 (or 100%), the latter arising when all sequences are unique. Thus, higher values indicate lower levels of commonality (or greater uniqueness).

The two measures outlined above may be expressed more generally as

$$MSF = \frac{n}{u+s} \qquad \text{and} \qquad SUI = \frac{u}{n} \qquad \text{where}$$

n = the total number of individual subjects represented in the sequence set

u = the number of unique sequences in the set

s = the number of nonunique sequences in the set

While trivial, these measures provide convenient and readily interpretable quantitative indications of the degree of commonality which is present in a given set of sequences. They may be viewed as representing the level of individuality, diversity or heterogeneity of experience which exists across all the subjects in the dataset.

Returning to the second concept introduced above (that of complexity) the complexity characteristics of a set of sequences may be summarised by calculating the average number of state transitions experienced by subjects i.e.

$$\frac{\sum_s (t_i \times f_i)}{n} \qquad \text{where}$$

s = the number of sequences in the set

t_i = the number of transitions present in the ith sequence (i = {1, 2, 3... ...s})

f_i = the number of individuals for whom the ith sequence is observed

n = the total number of individual subjects represented in the sequence set

The above quantity, which is termed hereafter the *mean transition density* (MTD), may be thought of as a measure of the instability or fluidity of subjects' experiences over time. For the sequences of Appendix 5, the observed value is 2.38.

Values of the measures defined above are shown for all three sets of sequences in Table 14.6.1 *(next page).*

TABLE 14.6.1: Summary values for measures representing characteristics of sequence sets.

measure	value for SOCIAL POSITION (Appendix 5)	value for DAMPNESS (Appendix 6)	value for AIR POLLUTION (Appendix 7)
MSF	1.36	2.24	3.03
SUI	0.66	0.41	0.29
MTD	2.38	2.04	1.98

***

This completes the presentation of results which relate to subjects' detailed trajectories of experience over time. The next chapter includes material relating to occupational exposures.

CHAPTER 15: RESULTS (V) - MEASURES OF OCCUPATIONAL EXPOSURE; PROPERTIES AND ASSOCIATIONS

15.1 Summary properties of the measures of occupational risk exposure

The derivation of three simple measures of exposure to occupational hazards was described in *Methods* Section 8.2. Each of these is a binary quantity representing whether the subject's lifetime exposure to a specific hazard (e.g. physically arduous work) was deemed 'low' or 'high'[44]. The prevalence of high exposure to each hazard is shown in Table 15.1.1.

TABLE 15.1.1: Observed prevalence values for high lifetime exposure to three specific occupational hazards. Cell content is number of positives (individuals for whom high exposure is observed), followed by % of positives, followed by total number of cases.

hazard type: exposure to...	**MALES *n* positive / % prevalence (total *n*)**	**FEMALES *n* positive / % prevalence (total *n*)**	**ALL *n* positive / % prevalence (total *n*)**	***p***
fumes / dusts	62 / 44.9 (138)	10 / 6.5 (154)	72 / 24.7 (292)	<0.001
physically arduous work	59 / 42.4 (139)	14 / 9.1 (154)	73 / 24.9 (293)	<0.001
demand / control stress	26 / 19.0 (137)	38 / 24.7 (154)	64 / 22.0 (291)	0.26

NOTE: Rightmost column shows the *p* value returned by a Fisher's exact test; small values indicate rejection of the null hypothesis of no association between sex and exposure to hazard.

15.2 Associations (i): occupational hazards and health

Section 8.3 justified the investigation of three sets of associations involving the measures of occupational exposure (shown in Table 15.1.1), in the interests of providing some semi-formal insight into whether subjects' experience of work-related hazards might exert a confounding influence on the main relationships of interest in the study. The first such set of associations consisted of those between the markers of occupational exposure and the main health outcomes featured in the study. Table 15.2.1 *(next page)* presents results from investigations of this first set of associations.

[44] These categories are defined in terms of the duration of the exposure, not its intensity (see Section 8.2).

TABLE 15.2.1: Summary of associations between the main health measures and three indicators of exposure to occupational hazards. Cell content is the *p* value returned by the appropriate statistical test (see Notes).

health measure	**exposure to FUMES/ DUSTS** ***p***	**exposure to ARDUOUS WORK** ***p***	**exposure to DEMAND STRESS** ***p***
physiological [note 1]			
systolic BP	0.87	0.72	0.24
diastolic BP	0.53	0.88	0.21
standardised FEV1	0.04	< 0.01	0.52
clinical [note 2]			
heart disease	0.08	0.48	1.00
lung disease	0.38	1.00	<0.001
stroke	0.15	0.74	0.07
high blood pressure	0.21	0.88	0.33
medication use [note 2]			
anti-hypertensive	0.77	0.55	0.06
bronchodilator	0.50	0.51	<0.001

NOTE 1: Results shown are *p* values returned by a Mann-Whitney test for equivalence of values between exposure subgroups (low *vs.* high).

NOTE 2: Results shown are *p* values returned by a Fisher's exact test.

15.3 Associations (ii): mutual relationships among the occupational hazard indicators

The second set of associations outlined in Section 8.3 consisted of relationships among the three binary indicators of occupational exposure. Tables 15.3.1 to 15.3.3 illustrate these associations via 2 X 2 contingency tables, each table being accompanied by the result of a Fisher's exact test.

TABLE 15.3.1: Contingency table of exposure to fumes / dusts (rows) with exposure to arduous work (columns). Cell content is number of subjects, followed by row percentage and column percentage.

	arduous work: LOW	arduous work: HIGH
fumes / dusts: LOW	191 (86.8 / 87.2)	29 (13.2 / 39.7)
fumes / dusts: HIGH	28 (38.9 / 12.8)	44 (61.1 / 60.3)

NOTE: Fisher's exact test returns $p < 0.0001$.

TABLE 15.3.2: Contingency table of exposure to fumes / dusts (rows) with exposure to demand / control stress (columns). Cell content is number of subjects, followed by row percentage and column percentage.

	demand stress: LOW	demand stress: HIGH
fumes / dusts: LOW	176 (80.4 / 77.5)	43 (19.6 / 67.2)
fumes / dusts: HIGH	51 (70.8 / 22.5)	21 (29.2 / 32.8)

NOTE: Fisher's exact test returns $p = 0.10$.

TABLE 15.3.3: Contingency table of exposure to arduous work (rows) with exposure to demand / control stress (columns). Cell content is number of subjects, followed by row percentage and column percentage.

	demand stress: LOW	demand stress: HIGH
arduous work: LOW	177 (81.2 / 78.0)	41 (18.8 / 64.1)
arduous work: HIGH	50 (68.5 / 22.0)	23 (31.5 / 35.9)

NOTE: Fisher's exact test returns $p = 0.03$.

15.4 Associations (iii): occupational hazards and smoking

The final set of associations outlined in Section 8.3 comprised relationships between the dichotomous indicators of work-related risk and subjects' smoking behaviour, the latter represented by a binary variable contrasting current or former smokers against those who have never smoked (see Section 5.3.5). Tables 15.4.1 to 15.4.3 illustrate these relationships in tabular form, together with the results of appropriate hypothesis tests.

TABLE 15.4.1: Contingency table of exposure to fumes / dusts (rows) with smoking status (columns). Cell content is number of subjects, followed by row percentage and column percentage.

	never smoked	current / former smoker
fumes / dusts: LOW	82 (37.3 / 90.1)	138 (62.7 / 68.7)
fumes / dusts: HIGH	9 (12.5 / 9.9)	63 (87.5 / 31.3)

NOTE: Fisher's exact test returns $p < 0.0001$.

TABLE 15.4.2: Contingency table of exposure to arduous work (rows) with smoking status (columns). Cell content is number of subjects, followed by row percentage and column percentage.

	never smoked	current / former smoker
arduous work: LOW	80 (36.4 / 87.9)	140 (63.6 / 69.3)
arduous work: HIGH	11 (15.1 / 12.1)	62 (84.9 / 30.7)

NOTE: Fisher's exact test returns $p < 0.0001$.

TABLE 15.4.3: Contingency table of exposure to demand / control stress (rows) with smoking status (columns). Cell content is number of subjects, followed by row percentage and column percentage.

	never smoked	current / former smoker
demand stress: LOW	74 (32.6 / 81.3)	153 (67.4 / 76.5)
demand stress: HIGH	17 (26.6 / 18.7)	47 (73.4 / 23.5)

NOTE: Fisher's exact test returns $p = 0.45$

***

With this, presentation of the study's results is complete. Discussion begins in the next chapter.

CHAPTER 16: DISCUSSION (I) – ASSESSING THE CONCEPTUAL MODEL

16.1 Introduction

The original motivation behind this study was the desire to test a specific model representing one possible pathway via which the near-ubiquitous phenomenon of socially-patterned health might be explained. The possibility that housing conditions may mediate relationships between socioeconomic position and health is widely recognised (see Section 3.1). Moreover, it is intuitively attractive, given the substantial body of research evidence which has linked the residential environment to a range of health outcomes (although the present thesis has adopted a rather sceptical view of the value of some of that evidence). The notion that housing may act as a mediating factor in the social patterning of health is essentially one expression of the 'materialist' explanation for the phenomenon which was advanced in the Black Report. The possible role of housing in a materialist framework of health inequality was recognised in another seminal document in the literature devoted to social inequality in health:-

> "The structuralist / materialist explanation emphasizes the role of the external environment: the conditions under which people live and work... ...Inequalities in health in this context would come about because lower social groups are exposed to a more unhealthy environment. They do more dangerous work, *have poorer housing*, and have fewer resources available..." [emphasis added] (Whitehead 1992, p. 316)

Thus, the basic postulate of this study is one which has long been recognised as plausible. Consequently, the study was not intended to be innovative in a conceptual sense: it proposed no radically new mechanism or process to explain the health inequality phenomenon. What was novel was the fine degree of temporal detail with which both SEP and the putative mediating factor (aspects of the residential environment) were represented. Indeed, the availability of such rich data was one of the main motivating factors behind the study. Broadly speaking, the expectation was that the extremely detailed information provided by the Boyd Orr dataset would permit the associations of interest to be estimated on the basis of more precise measurements than have generally been used to date in health inequality research.

As the results presented in Chapter 13 show, relatively little evidence which was supportive of the postulated model was observed in the study. Results relating to each of the three

components of the model are now discussed in turn, beginning with the end-to-end relationship between social position and health.

16.2 Associations (i): social position and cardio-respiratory health

16.2.1 Summary of associations

The results presented in Chapter 13 (and summarised in Section 13.18) include some evidence for relationships between SEP and the measures of cardio-respiratory health which were used in the study. Specifically, higher levels of accumulated disadvantage were found to be significantly associated with poorer lung function (Section 13.2), and with increased odds of reporting stroke (Section 13.4). The multiple regression model reported in Section 13.2 estimates that for each additional year spent in disadvantage, standardised FEV_1 is reduced by 0.11 of one percent. Thus, a decade in disadvantage appears to be related to a reduction in FEV_1 of a little over one percent. The logistic model reported in Section 13.4 indicates that an additional year of exposure to disadvantage increases the odds of stroke by a factor of around 1.06. The estimated effect (i.e. the change in odds) of an additional decade in disadvantage is given by $1.06^{10} = 1.87$.

The detection of positive relationships between social disadvantage and, respectively, lung function and stroke is consistent with existing knowledge. For example, a review by Hegewald & Crapo concluded:-

> "There is a significant negative correlation between lung function (primarily FEV_1 and FVC) and SES. This relationship exists even after adjusting for smoking status, occupational exposures, and race." (Hegewald & Crapo 2007, p. 1608)

Similarly, Drever and Whitehead have highlighted class-related differentials in the experience of stroke, reporting "a pronounced gap in mortality between manual and non-manual classes for stroke". (Drever & Whitehead 1997, p. 100). It may therefore be concluded that the specific effects observed in this study are plausible, and to some extent expected.

In addition to these effects involving cumulative exposure to disadvantage, a significant relationship between the time-dependent (clustered) representation of social position and the experience of stroke was observed (Section 13.11). This finding is based on the result of a

hypothesis test (Freeman-Halton), and its interpretation is problematical. The observed prevalence of stroke in social Cluster 2 subjects (those who experienced persistent manual status; see Appendix 2) was 5 / 43 (= 11.6%), while the corresponding value for Cluster 1 (near-persistent non-manual status) was 1 / 94 (= 1.1%). Although the expected difference between the most disadvantaged individuals (Cluster 2) and those with minimal exposure to disadvantage (Cluster 1) was observed, the prevalence of stroke in four other clusters was actually zero, while that in two further groups exceeded the rate observed in Cluster 2. Thus, the expectation that the most and least disadvantaged groups would record, respectively, the 'worst' and 'best' stroke experiences in the sample was not rigidly fulfilled.

The three effects discussed above represent the only statistically significant associations involving SEP and health which were observed among the large number examined. Of nine individual associations involving cumulative disadvantage, only the two highlighted above were judged significant at the conventional 5% level, while a third (that between disadvantage and bronchodilator use) was deemed marginally significant ($p = 0.06$)[45]. Nine further associations between time-dependent social position and cardio-respiratory health were assessed; of these, only that involving stroke was judged statistically significant. When considered *en bloc*, the study's results do not suggest the presence of convincing relationships between social position and the specific areas of health examined by the study (that is, disorders of the cardiovascular and respiratory systems). This is an unexpected finding, because social disadvantage has been specifically demonstrated to show associations with both of these classes of disease. A review by Kaplan & Keil concluded "The evidence appears to support the argument that SES is an important factor in the etiology and progression of cardiovascular disease." (Kaplan & Keil 1993, p. 1973). Similarly, evidence for associations between low SEP and poor respiratory health has been summarised by Strachan, who asserts:-

> "It has long been recognised that there is a strong association between poor socioeconomic status and both mortality and general practice consultations for adult respiratory disease in Britain." (Strachan 1997, p. 111)

The largely uncontested view that social disadvantage is linked to both cardiovascular and respiratory health makes it appropriate to consider possible explanations for the tenuous nature of that relationship as it is observed in the present study. Five such explanations suggest themselves, and are now discussed.

[45] See the footnote to Section 13.6 for clarification of the concept of 'marginal' significance.

16.2.2 Explaining the absence of association

A first possible explanation for the absence of convincing associations is that the findings reflect actual uncertainties in the dataset itself. While there is no reason to question the accuracy of the health outcome measures, the representation of social position in the dataset is based on subjects' memories of experiences which occurred up to several decades in the past. It might be argued that these memories are in some cases unreliable, with the result that the derived measures of social location used in the study are founded on potentially inaccurate recollections of subjects' true experiences. However, the accuracy of the information recalled via the lifegrid method has been considered at an earlier stage (see Section 5.2), and the quality of data so obtained appears to be acceptable. Consequently, this explanation may reasonably be discounted.

A second possibility is that the available sample size is simply too small to permit the detection of effects which may be of modest magnitude. The dataset contains details of 294 individuals, and the *effective* sample size (i.e. the number of subjects actually usable in a specific analysis) is in many cases smaller than this due to the effects of missingness (see Section 9.2). Sample size, and the related concept of statistical power, is a fundamental concern in many studies. Comments on this topic by Altman relate to experimental contexts, but are equally applicable to observational studies such as the present one:-

> "The medical literature contains many trials that were far too small to have a good chance of detecting clinically worthwhile differences between the treatments being investigated... ...Unless the true treatment effect is large, small trials can yield a statistically significant result only if, by chance, the observed difference in the sample is much larger than the real difference." (Altman 1991, p. 455)

Altman's comments on the possible limitations imposed by sample size lead directly to a third potential explanation for the present study's findings. It is conceivable that the individuals featured in the dataset are in some way atypical of the wider population from which they are drawn, with the result that associations which might be present in that population are diluted or absent due to peculiarities of the sample. This is in effect the reverse of the scenario identified by Altman: if the association present in a sample is *smaller* than the true relationship in the population, statistically significant results are less likely to be observed. This possibility was introduced in Section 5.2, which acknowledged that the sample incorporates a conservative bias, in that the most disadvantaged individuals in the

original Boyd Orr survey were disproportionately lost to follow-up through death. Moreover, non-responders to the request for interview in old age were found to be more disadvantaged as children than those who agreed to be interviewed. Thus, the sample is characterised by levels of disadvantage which are not representative of the notional population from which it is drawn. Furthermore, it is also arguable that the individuals in the dataset are atypical of their population in terms of health. As highlighted in Section 2.1, subjects were aged between 63 and 78 years when interviewed, and the fact that the sample is composed entirely of people who were still alive at such relatively advanced ages suggests a further possible source of bias. Overall, it is conceivable that the limited associations observed in the study may reflect the lower experience of disadvantage and the superior level of health which characterise the sample. This hypothesis cannot be tested, and must remain conjectural.

A fourth potential explanation is that the actual measures of social position which were constructed from respondents' original recollections (i.e. the cumulative disadvantage total of Section 6.2, and the time-dependent representation of Section 6.3) are flawed. That is, while subjects' original recollections are essentially correct, the translation of these recalled experiences into new derived variables may have involved unwarrantable assumptions, and thus introduced a degree of error. However, the basis on which the new variables were derived has been explained and justified in considerable detail. The use of clustering to create a grouping scheme for use in an inferential (as distinct from an exploratory) role is arguably the aspect of the methodology which is most vulnerable to criticism. The justification for this approach was given in Section 6.3.3, and precedent for its adoption exists. For example, McVicar and Anyadike-Danes (2002) used clustering (based on distance values obtained via optimal matching) to create grouped trajectories representing the educational or employment status of young people during the transition from school to work. These grouped trajectories were then used as the outcome in a formal statistical model which sought to identify the determinants of group membership.

A final potential explanation for the limited associations observed between SEP and health relates to the possible confounding influence of occupational exposures. Chapter 8 described the derivation of three indicators of work-related risk, and Chapter 15 reported the results of investigations involving these measures. A first set of investigations assessed relationships between the occupational indicators and the main health outcomes. Of 21 individual associations examined, four (all related to respiratory health) were significant at the

conventional 5% level (see Table 15.2.1). Lung function (standardised FEV_1) was associated with exposure to fumes and with physically arduous work, while the experience of lung disease was linked to job demand / control stress. The latter was also associated with use of bronchodilator medication. On this basis, a possible confounding effect of occupational risk on the associations between SEP and respiratory heath which were investigated by the study is identified.

However, interpretation of the findings presented in table 15.2.1 is problematical, for two reasons. First, the additional investigations reported in Chapter 15 indicated that the occupational risk factors were themselves inter-related. Exposure to fumes and the experience of arduous work were significantly associated (Table 15.3.1), and the latter also exhibited a significant relationship with demand / control stress (Table 15.3.3). Consequently, any respective effects of the three occupational exposures on the main associations of interest in the study are to some extent inseparable. Second, two of the occupational hazards (exposure to fumes, and arduous work) exhibited very highly significant associations ($p < 0.0001$) with subjects' smoking status (Tables 15.4.1 and 15.4.2). As a result, disentangling the respective influences of work-related risks and of smoking on the associations of interest would be challenging. A series of stratified analyses would provide some clarification, but this was not considered feasible: the attenuation of an already-small sample by stratification would present significant analytical difficulties (see Section 8.3).

Although it was not possible to adjust for the potential confounding influence of occupational exposures, it is considered extremely unlikely that the observed lack of association between SEP and respiratory health reflects confounding effects. The failure to control for work-related exposures would, if anything, lead to overestimation of the relationships involved. Since these relationships are tenuous to begin with, further attenuation resulting from controlling for occupational risk factors (were it possible) would in all likelihood simply intensify the challenge of explaining the lack of convincing association.

In summary, five potential explanations for the relative weakness of the observed associations between social position and health have been considered. It is argued that the first of these (inaccuracies in the dataset) may be discounted. Two further possibilities (lack of statistical power due to small sample size, and non-representativeness of the sample)

cannot be tested, and must remain conjectural. A fourth explanation focuses on methodological inadequacies associated with the derivation of variables, but the processes involved are justified at some length in the relevant *Methods* sections. A final possibility (failure to control for work-related exposures) is viewed as unconvincing, because the expected effect of such a failure would be overestimation of the associations of interest.

***

The absence of a robust overall relationship between SEP and cardio-respiratory health carries obvious implications for the plausibility of the conceptual model investigated by the study (i.e. that housing conditions mediate the link between disadvantage and health). The model is a unified whole: each of its three component associations (i.e. SEP with health, SEP with residential conditions, and housing conditions with health) is necessary, and none is individually sufficient for the model to be considered valid. The failure to detect robust relationships in the end-to-end link between social position and health indicates that the overall model is not supported by these data.

Despite this, the two remaining sets of associations postulated in the model remain of interest. While many studies have examined aspects of these associations (in particular, relationships between the residential environment and health), few have done so using measures which represent the exposures of interest over an extended time period to a fine degree of temporal granularity. The present study's recording of exposure at the level of the individual year is unusual, and consideration of the remaining associations is therefore warranted even although the validity of the model as a whole is discounted by the failure to detect a robust end-to-end association across the posited causal chain. Discussion continues by considering the study's findings in relation to the rightmost association in Figure 4.2.1 (that between housing conditions and health).

16.3 Associations (ii): exposure to residential hazards and cardio-respiratory health

As reported in Section 13.3, nine individual associations between the cumulative experience of residential risk and the physiological health measures were assessed, and none was judged statistically significant at the conventional 5% level. A further twelve associations between accumulated residential hazard exposure and the clinical variables were examined (Section 13.5). Of these, only the link between cumulative exposure to dampness and lung disease

was deemed significant. The logistic model for this association estimates that an additional year of exposure to dampness increases the odds of lung disease by a factor of 1.04; an additional decade of dampness exposure raises the odds of lung disease by a factor of 1.54. Six associations between cumulative residential hazard exposure and use of medications were assessed, none of these emerging as statistically significant (Section 13.7). Finally, a further eighteen individual associations between the clustered representations of housing risk and the cardio-respiratory health outcomes were evaluated. A significant relationship between dampness and systolic blood pressure was identified (Section 13.10). However, interpretation of this result is problematical. Table 13.10.2 indicates that systolic blood pressure was lower (i.e. 'better') among individuals in dampness Cluster 3 (who experienced substantial periods of exposure; see Appendix 3) than in the zero-exposure Cluster 1. Finally, a marginally significant ($p = 0.06$) link between dampness and reported lung disease was detected (Section 13.12). None of the remaining associations was statistically significant.

Overall, the evidence for the existence of convincing relationships in this component of the study's postulated conceptual model is extremely weak. Among possible explanations for this finding, two are similar to those discussed in Section 16.2.2: inaccuracies in the dataset (i.e. subjects' defective recollections of their housing histories), and a lack of statistical power associated with the small available sample size. While the importance of the latter must remain conjectural, the issue of accuracy of recollection merits further consideration, particularly in connection with the experience of dampness. The identification of residential damp was based on memories of 'black mould or other signs of damp' (see the extract from the *Users Guide to the Dataset* which appears in Section 5.3.2). However, as discussed in Section 3.3.3, it has been demonstrated that visible manifestations of mould may be an imprecise indicator of actual mould exposure. Moreover, it is possible that any uncertainties or imprecision in the reporting of mould might be amplified, perhaps substantially, when the dwellings involved were occupied many years in the past. Consequently, even if the subject's recollections of the presence of mould were broadly accurate (in the sense that *some* level of mould was or was not present in her / his *N*th dwelling), this may not necessarily provide a correct representation of the extent to which s/he actually experienced dampness in the home. The measures of exposure to dampness which were used in the study are based entirely on these recollections, and the findings are therefore highly vulnerable to any imprecision in subjects' reporting of the manifestations of dampness. In the absence of any way of validating the reports of residential conditions which feature in the dataset, it is

impossible to determine whether the observed results reflect any influence of defective reporting.

A third possible explanation for the observed lack of association between residential hazards and health is that inaccuracies were introduced during derivation of the representations of hazard exposure which were used in analysis. As was described in Chapter 7, the creation of residential histories and associated measures of hazard exposure involved extensive manipulation of the original data, underpinned by certain assumptions. While these assumptions are justified at length in the relevant areas of Chapter 7, the possibility that the processes involved introduced biases or inaccuracies cannot be completely discounted. However, it is unlikely that any such biases would be sufficiently gross as to be solely responsible for the overwhelmingly negative findings observed.

***

Thus far, consideration of the study's findings has concluded that two of the three component associations of the model of Figure 4.2.1 are not convincingly demonstrated. From this, it might be argued that further consideration of the study's results in relation to the model is of little value. The weakness of the end-to-end relationship between social location and health, coupled with the almost complete absence of any meaningful association between housing conditions and health, leads inexorably to the conclusion that the general model (and hence its three realisations in Models A, B and C) is not supported by these data. However, in the interests of completeness, results relating to the remaining component of the general model (the postulated link between social position and exposure to adverse residential conditions) are now briefly discussed.

16.4 Associations (iii): social disadvantage and exposure to residential hazards

The results presented in Section 13.16 indicate that cumulative social disadvantage is significantly predictive of both exposure to air pollution and the individual's total hazard load (Table 13.16.1). The direction of effect is, in both cases, that which would have been anticipated: increased disadvantage is associated with greater experience of both hazards. The effect sizes are modest, the regression models involved estimating that each additional year spent in disadvantage is associated with an extra 0.15 year of exposure to air pollution, and an additional 0.23 year increase in total hazard load. In addition to these associations, a

marginally significant positive relationship between disadvantage and the experience of residential dampness was observed, although the point estimate of the effect size is very small. In contrast to the above associations (which involve accumulated social disadvantage), the grouped representation of social position was found not to be significantly associated with the time-dependent experience of either dampness or air pollution (Section 13.17).

Although modest, the significant associations observed are consistent with the intuitively plausible expectation that individuals holding lower SEP will be exposed to a residential environment of poorer quality than that enjoyed by those in higher socioeconomic strata. While expressions of this expectation are common (see for example the quotations from Hopton *et al*. and Shaw *et al*. which were presented in the opening paragraph of Section 3.1), formal testing of the hypothesis that social disadvantage is associated with exposure to specific housing hazards appears to be relatively rare in the health inequality literature. The reasons for this are understandable. Outwith the social housing sector, housing is a market commodity and may therefore confidently be expected to display strict adherence to the laws of market economics. Consequently, it is entirely reasonable to expect that poor people will in general be restricted to the less desirable (and, by implication, more hazardous) end of the spectrum of housing quality.

***

At this point, comment has been made on results which relate respectively to the three individual components of the conceptual model tested in the study. Some general remarks relating to the overall validity of the model are now presented.

16.5 Assessing the model: concluding comments

The study's overall findings in relation to the model proposed in Section 4.2 are decisively negative. It is not possible to argue, on the basis of the observed results, that the posited mediating effect of residential conditions on the relationship between SEP and health is convincingly supported. Interpretation of the negative findings is hindered by the unusual nature of the representations of both the putative risk factor (i.e. SEP) and the hypothesised mediating element (residential conditions). Few other studies have represented both lifetime SEP and cumulative housing exposures to the level of the individual year; consequently,

there is little opportunity to compare results from this study with those of analogous earlier studies which have employed similarly 'fine-grained' representations of social and / or residential exposures. Such a comparison might have illuminated the findings, suggesting explanations for the overwhelmingly negative results beyond those which are considered in the earlier sections of the present chapter.

The failure to demonstrate the validity of the model effectively dictates an unequivocally negative response to the study's second research question (i.e. whether specific patterns of exposure to disadvantage and / or poor housing conditions predict cardio-respiratory health in old age). The study clearly did not identify robust relationships between such exposures and the specific indicators of cardio-respiratory health which were investigated. Expressed more formally, there is insufficient evidence in these data to support the hypothesis that differentials in SEP across adult life lead to variations in cardio-respiratory health in old age, via a mediating influence of the residential environment. As was discussed in Section 3.1, the interlinked nature of the three factors involved (i.e. SEP, housing conditions and health) is widely recognised, and the difficulty of identifying the direction of the associations which link them is acknowledged. In planning the study, the expectation was that some elucidation of the relationships among these factors would be achieved. While this proved not to be the case (in the sense that the finding were decidedly negative), it is an accepted part of scientific thinking that such findings – which may be viewed as 'the demonstration of a null' – are far from valueless, provided that they do not result from identifiable flaws in the research process which generated them. While issues relating to negative findings (and the associated phenomenon of publication bias [Dickersin, 1990]) are perhaps most commonly discussed in relation to clinical trials (Altman, 1991), the topic is also relevant to purely observational studies such as the present one. That the study's findings are negative in relation to this research question is in itself of some interest, in that it provokes further thought as to whether the conceptual model underlying the study is plausible.

***

As outlined in Section 4.1, the model raises questions which extend beyond the nature of the associations linking the hypothesised determinant of health (SEP) and the postulated mediating factor (housing conditions) with each other, and with health. A second focus of interest is how exposure to the determinant and the mediator behave over time; specifically,

over adult life. This topic, which forms the basis of the study's first research question, is treated in the next chapter.

CHAPTER 17: DISCUSSION (II) – VARIATION IN SOCIAL AND RESIDENTIAL EXPOSURES OVER TIME

17.1 Introduction

While one objective of the study was to assess evidence relating to a postulated mediating effect of housing conditions in the SEP / health relationship, the examination of variation over time in both social position and exposure to residential hazards is a natural complement to testing the validity of the model itself. Both an individual's SEP and the characteristics of her / his residential environment are subject to change with time, and it is reasonable to postulate that the nature of any associations between these factors, and their respective relationships with health, will potentially be influenced by the form which such change takes. Therefore, a comprehensive investigation of the model may legitimately extend to consideration of how its two predictive factors behave over time.

A distinguishing feature of this study, emphasised throughout this thesis, is the unusually detailed representations of both SEP and exposure to residential risks which were available for use in analysis. With regard to the former, the availability of year-on-year information on occupational class over a period of several decades is in marked contrast to many studies in the health inequality field. While a variety of study designs have been adopted to investigate the social patterning of health, one commonly-used approach is to sample the subject's SEP at a small number of widely spaced time points (typically three), and use this information to construct either a measure of accumulated lifetime social disadvantage or (more rarely) a set of explicit trajectories of SEP over the lifecourse. The former approach is exemplified in the study of Davey Smith *et al*. (1997) which was described in Section 6.2.1. Other studies adhering to the same general pattern (i.e. estimating cumulative disadvantage from a three-point sampling scheme) include those reported by Lynch *et al.* (1997), Heslop *et al.* (2001), Pensola & Martikainen (2003), Adams *et al*. (2004), Naess *et al*. (2004a) and Singh-Manoux *et al*. (2004)[46]. The popularity of such sampling regimes is attested to by Ljung & Hallqvist:-

> "The measure of accumulated adverse socioeconomic positions is in most studies rather crude and is based on a combination of father's socioeconomic position

[46] In the study of Pensola & Martikainen, the calculation of accumulated disadvantage on the basis of three sampling points represented only a small part of the study, which was concerned mainly with a two-stage representation of lifetime socioeconomic position (based on SEP at childhood and in adulthood).

> (childhood), highest educational attainment or first occupation (young adulthood) and current occupation-based socioeconomic position (adulthood)." (Ljung & Hallqvist 2006, p. 1080)

This statement is noteworthy in two respects. In addition to confirming the widespread use of three-point sampling of lifetime SEP, it acknowledges the limitations of measures derived in this way (characterising them as 'rather crude'). The latter idea is developed at considerable length throughout the *Discussion* chapters of this thesis.

On occasions, more than three sampling occasions are used; for example, Power and colleagues (1999) and Naess *et al*. (2006) constructed measures of lifetime disadvantage based on the subject's SEP at four time points. In contrast to such uses of cumulative representations, instances of SEP collected at a small number of points being exploited to create explicit trajectories of social experience include studies by Adams *et al*. (2004), Hallqvist *et al*. (2004) and Naess *et al*. (2006)[47]. In the latter study, the researchers' determination of subjects' SEP (defined as advantaged or disadvantaged) at four time points permitted recognition of $2^4 = 16$ distinct trajectories. Comments by these authors on the advantages and disadvantages of trajectory-based approaches (relative to accumulated measures) are relevant to much of the discussion which follows, and merit quotation:-

> "The cumulative model has some potential when explaining distribution of various chronic diseases in western populations, but it fails to take into account the temporal sequence of causal influences... ...They *[several authors are cited]* suggested a pathway approach which specifies more in detail how risk may vary between different trajectories. Such an approach may enable us to identify, in detail, stages in the life course that has *[sic]* a detrimental influence. Furthermore, it may provide deeper exploration of trajectories that have similarities in their influence on risk of ill health. Few studies have investigated this pathway approach empirically. Due to increased detail in the model, it demands large data to conduct." (Naess *et al*. 2006, p. 330)

Whether expressed as measures of cumulative disadvantage, or as systems of actual trajectories, representations of lifetime disadvantage which are based on establishing the individual's SEP at a small number of points are (as demonstrated above) common in health inequality research. Much of the corpus of current knowledge relating to social inequalities in health has been obtained from studies which have employed such representations. However, these 'sparse' sampling schemes suffer from an obvious potential drawback. Where the person's SEP is captured at (say) childhood, early adulthood and middle age, no information is available on what happens between these sampling points. This kind of sparse

[47] The studies of Adams *et al*. (2004) and Naess *et al*. (2006) created both a cumulative measure of disadvantage and actual trajectories of SEP.

sampling regime has been discussed by Pickles & De Stavola (2007), who express the scheme used in the influential study of Davey Smith *et al.* (1997) mathematically as

$$\sum_{t=1}^{3} x(t)\Delta(t),$$

where *x(t)* is a binary indicator representing exposure to disadvantage (coded one) *vs.* freedom from such exposure (coded zero) at point *t;* $\Delta(t)$ is the period of exposure associated with *t*; and all $\Delta(t)$ are assumed to be of equal length. The potential for imprecision inherent in such a scheme is substantial, and easily illustrated. Consider the case in which an individual is deemed exposed to disadvantage at time t_i (representing, say, the point at which the person entered the labour market), but not exposed at the next sampling point t_j (when s/he is recruited into a study). The transition from exposed to unexposed could theoretically take place at any time point from just after t_i to just before t_j. Thus, retaining the notation of Pickles and De Stavola, the length of the exposure interval $\Delta(t)$, although constrained, is in fact indeterminate. Consequently, the summation of exposure over all $\Delta(t)$ may provide a biased or inaccurate estimate of the quantity which is really of interest (but unobtainable), namely the integration of point exposure levels over the entire period (say, *T1* to *T2*) i.e.

$$\int_{T1}^{T2} x(t)dt \qquad \text{(Pickles \& De Stavola, 2007)}$$

This limitation justifies the previously-cited characterisation of sparse sampling designs as 'rather crude' ((Ljung & Hallqvist, 2006).

By capturing subjects' exposure to social disadvantage more precisely than is the norm in inequality research, the present study permitted lifetime social experience to be examined in considerably more detail than is generally possible in inequality studies. The measure of accumulated disadvantage used here is closer to the ideal form identified by Pickles & De Stavola (expressed in the integral presented above) than is often the case. Similarly, the sequences or trajectories of social position which were constructed arguably provide a more accurate representation of the individual's social path through life than the three-point trajectories featured in previous studies. Consequently, this project offered the opportunity to assess how effectively the sparse sampling schemes which have often been used in inequality research can represent the reality of an individual's actual social experience. The same advantage extends to the study's representation of exposure to residential hazards:

these were examined in far greater detail than is often the case in studies of housing conditions and health.

In summary, two motives are suggested for examining the detailed, year-on-year representations of social and residential experience which were created for the study. First, as discussed at the start of this section, investigating time-related variation in these factors is a natural part of any overall assessment of the conceptual model presented in Section 4.2. Second, consideration of how social position in particular varies over the lifecourse provides insights into how effectively the sparse sampling schemes which have often been used in inequality research can represent the individual's social experience. The latter topic - the adequacy of the classic sampling scheme which has been the foundation of much seminal research in health inequality - is a theme which is returned to repeatedly in the material which follows.

***

Discussion now proceeds to consider the study's findings relating to subjects' lifetime social experiences.

17.2 Social experiences over the lifecourse

17.2.1 Accumulated disadvantage

The distribution of accumulated disadvantage was illustrated in Figure 11.2.1, and summary statistics for the measure were presented in Table 11.2.1. Two features of interest are evident. First, the experience of disadvantage appears to differ between the sexes; this is suggested visually by the histograms of Figure 11.2.1, and is formally confirmed by the significant Mann-Whitney test result reported in the associated table. Second, despite this difference, the distribution of the measure exhibits a distinct bimodal form which (although less pronounced for men) is common to both sexes. One modal point occurs towards the lower end of the distribution (but not at the zero point), while a second is located at the upper extreme.

The observed between-sex difference presents problems of interpretation. As explained in Section 6.1.3, the method used to determine subjects' SEP (and hence their exposure to

disadvantage) at the ages of interest differed between men and women. Specifically, where no occupational social class was recorded for a married woman at age *Y* years, the main lifetime class of her husband was assigned. Thus, it may be argued that consideration of between-sex differences in the experience of disadvantage is slightly artificial, because the basis of calculation is not consistent. The general effect of partially imputing to women the social class of their husbands is to make the disadvantage experience of females 'more similar to' that of males. Consequently, the between-sex differences evident in Figure 11.2.1 might be regarded as a diluted or vitiated representation of the 'true' difference in the respective disadvantage experiences of the sexes.

The manipulation applied to derive the disadvantage measure for women makes any comparison with other studies (which may not have applied a similar imputation process) unsafe. However, the distribution of cumulative disadvantage observed in this study for men may tentatively be compared with the corresponding distributions observed by other researchers. Such comparisons are subject to obvious caveats (such as different age ranges being covered, or samples being drawn from very specific populations). Mindful of such considerations, Table 17.2.1 shows the distribution of accumulated disadvantage observed for male subjects in two British health inequality studies which have been cited previously in this thesis: the influential study of Davey Smith *et al*. (1997), and that of Adams *et al*. (2004). Both of these captured subjects' social position (manual *vs.* non-manual) at three points, constructing a disadvantage measure which represented the number of occasions (minimum zero, maximum three) on which the person held manual status, and was consequently presumed to be disadvantaged. The table shows the proportion (percentage) of male subjects who fell into each of the four ordered disadvantage groups in each study.

TABLE 17.2.1: Number and percentage of male subjects assigned to lifetime disadvantage groups in studies by Davey Smith *et al*. (1997) and Adams *et al*. (2004). Cell content is number of individuals, followed by percentage of total (columns sum to 100%).

cumulative disadvantage (no. of occasions)	**Davey Smith *et al*. (1997) *n* (% of total)**	**Adams *et al*. (2004) *n* (% of total)**
0	956 (17.2)	39 (19.3)
1	1122 (20.2)	52 (25.7)
2	1170 (21.0)	42 (20.8)
3	2319 (41.7)	69 (34.2)

At the risk of labouring the point, direct comparison of the values in Table 17.2.1 with the distribution of cumulative disadvantage observed in the present study can only be tentative, due to the presence of potentially important differences. For example, Adams and colleagues examined the age range from birth to 50 years, in contrast to the period from 15

to 60 covered by this study. Nevertheless, it appears initially reasonable to argue that, in a very general sense, the distribution of disadvantage in these other studies is not drastically different from that shown in the left-hand panel of Figure 11.2.1. A pronounced 'spike' at the upper end of the distribution (i.e. the point signifying greatest disadvantage) is common to all three studies, as is a broadly uniform distribution of subjects across the remainder of the range. It might even be suggested that the data collected by Adams exhibit, in very crude form, the bimodality observed in the present study. This coarse resemblance between the results of the studies initially suggests two conclusions. The first of these relates only to the present study: the broad similarity of the distribution of disadvantage observed here to that found in previous research provides a modest measure of reassurance that there is nothing markedly distinctive or abnormal about the social experiences of the subjects who comprise the Boyd Orr dataset. This, in turn, supports the generalisation of this study's findings to a wider population.

A second conclusion has more extended applicability. The representations of disadvantage constructed by Davey Smith *et al*. and Adams *et al*. were based on sparse sampling conforming to the pattern discussed in Section 17.1. Comparison of these researchers' findings with those of the present study (which are derived from considerably more detailed data) initially suggests that sparse sampling has, in these two instances, been fairly successful at capturing the major features of subjects' lifetime social experiences. If the present study may be regarded as an informal standard or norm (by virtue of the highly detailed SEP information on which it is based), then the more economical sampling regimes of the other studies may be regarded as being, to some extent, validated.

However, this initial conclusion merits more detailed consideration. Returning to Table 17.2.1, the proportions of subjects deemed to be completely free from exposure to disadvantage (those in the uppermost row) are 17.2% (Davey Smith) and 19.3% (Adams). That is, around one-fifth of the men in each study are assumed (under the three-point sampling schemes used) to have experienced no social disadvantage. In the present study, the proportion of male individuals for whom no disadvantage is recorded is 2 / 133 (= 1.5%). While the practical significance of absolute freedom from disadvantage is debatable, it appears that the sparse sampling regimes which have featured in much health inequality research may be limited in their ability to detect this particular state. More generally, while coarse sampling of lifetime SEP at (say) three points may successfully capture the *general* nature of subjects' social experiences, it cannot detect more subtle features (such as the very

low prevalence of 'pure' freedom from disadvantage). Such features may, or may not, be of importance. Given its dominance in health inequality studies, and the extent to which it has shaped current knowledge, the adequacy of sparse sampling schemes is a topic with important implications for both the interpretation of previous studies and the conduct of future research.

***

Thus far, consideration of subjects' social experiences has been limited to their cumulative exposure to disadvantage. Discussion now shifts to the other perspective on lifetime SEP provided by the study, namely the detailed trajectories of social position whose construction was described in Section 6.1.

17.2.2 Social trajectories (i): patterns of transition

A convenient starting point for discussion of the social trajectories (which are presented in full in Appendix 5) is provided by the summary quantities introduced in *Results* section 14.6. One of these measures (the mean transition density [MTD]) was designed to quantify that property of the trajectories which was termed *complexity*. The MTD represents the average number of state transitions observed per individual in the dataset, and so provides an indication of how much instability or flux is present in subjects' social experiences. Over the 46-year period examined, the mean number of changes in social state experienced by subjects was as shown in Table 14.6.1 i.e. 2.38. Interpretation of this value is difficult because no recognised norms exist against which it may be compared. This also applies to the other quantities defined in Section 14.6, reflecting the exceptional nature of the data used here. Despite this, it is reasonable to make certain observations based on the values of these measures. One interpretation of the MTD value of *c.* 2.4 is that the adult lives of this sample were characterised by relatively little social flux (or, alternatively, by a reasonable degree of social stability). As shown earlier (Figure 14.4.2), the distribution of the number of social transition events was actually skewed, a small number of individuals recording 'turbulent' patterns of experience characterised by frequent changes in SEP.

The indication that individuals represented in the dataset experienced an average of (roughly) two and a half changes in social position during adult working life stimulates further consideration of the sparse sampling schemes typically used in inequality research.

One feature of such schemes, perhaps not immediately obvious, is that they impose limits on the number of social transitions which are assumed to occur. With the commonly-used three-point sampling regime, the maximum number of transitions allowed for is two (e.g. from state *A* [say, occupational class IV] at point 1 to state *B* [e.g. OSC II] at point 2, then back to state *A*). That this sampling approach permits a *maximum* of two transitions, while the *average* number observed in the present study exceeds this, suggests another possible limitation of infrequent sampling: it may not adequately cater for the level of flux or dynamism which is actually present in the social experiences of many people.

In addition to the absolute numbers of social transitions observed, their distribution by age is of interest. The location of social transition points by age in the present study was shown in Figure 14.4.1, which reveals that changes in social state tended to be concentrated in young adulthood. When viewing the figure it is helpful to recall that the substantial transition activity evident at age 18 is partly artefactual, reflecting the decision to impute the parental (father's) occupational class to the subject up until the age of 17 (see *Methods* Section 6.1.2). Because of this, changes in personal occupational class resulting from entering the labour market at (say) age 16 will not become manifest until age 18. However, even when this effect is disregarded, Figure 14.4.1 suggests that the early 20s (and, for men, the late teens) represent the period of greatest social flux. The same effect is evident in the position weight matrices of Figure 14.2.1a (males) and Figure 14.2.1b (females). The observation in this study of increased social stability from around age 30 onwards accords with findings obtained by sociologists; for example, it has been demonstrated that most men attain their destination social class by the age of 35 (Goldthorpe, 1980). The apparent concentration of social flux in early adulthood has further implications for sparse sampling regimes, suggesting that determining the individual's SEP more frequently in this 'turbulent' region might be desirable, while a more relaxed approach to capturing social position at multiple points in later life may be appropriate. In practice, sampling schedules (like many other aspects of epidemiological research) are usually dictated by practical considerations, such as what data are available, and the effort and resources required to collect them. However, in an 'ideal' study where the researcher had complete freedom to specify the sampling scheme, concentration on young adulthood might provide the optimum ratio of effort to results (the latter expressed in terms of the amount of useful information collected).

The specific nature of social change observed in the study was shown in detail in Table 14.4.1, which indicates that the most frequently observed types of transition observed for

men were the change from manual status to non-employment (21.8% of all transitions), and that from manual to non-manual status (18.2%). For women, the most common changes were those from manual to non-manual (40.1%) and from non-manual to manual (29.7%). These between-sex differences reflect the almost complete restriction of changes involving the Armed Forces state to men, and are also influenced by the attribution to married women of their husbands' occupational class. Although social mobility *per se* is not an interest of this study, Figure 14.2.1a/b suggests a general effect of upward mobility, the proportion of individuals in the manual state broadly decreasing with age while the corresponding proportion of those enjoying non-manual status increases.

17.2.3 Social trajectories (ii): the individuality of social experience

Further observations relating to the social experiences of the sample are suggested by the values of the other two summary quantities defined in Section 14.6: the mean sequence frequency (MSF) and the sequence uniqueness index (SUI). These quantities are measures of the property which was termed *commonality*, loosely defined as whether an observed trajectory (here, of SEP) is shared by multiple people, or is unique to a single individual. For these two measures, the social sequences of Appendix 5 yield the values reported earlier i.e. MSF = 1.36 and SUI = 0.66. In other words, each trajectory of social position observed in these data was shared by (on average) around 1.4 people, and two-thirds of the sample recorded patterns of social experience which were unique to them.

These values indicate a considerable (perhaps surprising) degree of individuality in the experience of socioeconomic position over adult life. This finding presented challenges in analysis, as discussed earlier in Section 6.3.1, which outlined the original motivation for condensing the raw sequences of Appendix 5 into higher-level groups in the interests of improved analytical tractability. However, beyond this essentially technical consideration, the substantial element of individuality evident in subjects' social trajectories has wider implications for research into health inequality. Specifically, it may be argued on the basis of the social sequence structure observed in this study that very general classifications of social experience which are commonly used in research (such as 'upwardly mobile' and 'downwardly mobile') are inadequate, in that such 'broad brush' concepts cannot fully represent the pronounced individuality of personal experience which is suggested by this study. To illustrate, reference is made to the previously-cited study by Adams and colleagues (2004) of the relationship between lifetime SEP and self-reported health at age 50

among *c.* 450 English adults. This identified "four socioeconomic trajectories (stable non-manual, upward, downward , and stable manual) over three time phases (birth to age 25, age 25 to age 50, and birth to age 50)." (Adams *et al.* 2004, p. 1028). However, the findings of the present study raise questions as to how effectively such a limited scheme represents the full individuality of personal experience, as expressed in the concept of commonality. For example, did *all* individuals in the 'stable manual' category of Adams *et al.* really experience manual status at each individual year throughout the period examined (i.e. birth to age 50)? Did *every* person in the 'upward' group genuinely undergo only a single social transition, moving from persistent manual status to a constant experience of the non-manual condition? The findings of the present study would suggest that this is unlikely, and that the (unknown) true experiences of Adams's subjects did not in fact strictly conform to such idealised and formally pure patterns.

The possibility that this study of Adams and colleagues may have involved some simplification of subjects' actual experiences may be generalised to highlight a further potential limitation of the classic sparse sampling design which has commonly been used in inequality research. A design in which SEP is collected at three stages, and is represented at each stage by two possible states (manual or non-manual), assumes that a *maximum* of eight possible distinct social trajectories exist[48]. Yet as the present study has shown, the actual number of distinct trajectories present in even a small sample of individuals may be much higher than this. Consequently, it may be argued that the classic design effectively forces individuals with unique social pathways through life into a system of crude categories which disguises much of that uniqueness. This in turn means that the relationships between social pathways and health, which are the core interest of health inequality studies, may be incorrectly estimated. However, a counter argument to this view is that although the number of unique social trajectories in a sample may be large, many of these trajectories might be viewed as very similar to each other, and may thus reasonably be regarded as representing a common experience. This argument is developed further in the next section.

17.2.4 Social trajectories (iii): unusual pathways and the problem of estimability

Thus far, discussion of the social trajectories has concentrated on the twin properties which have been labelled complexity (the frequency with which state transitions occur) and

[48] More generally, if some factor of interest (here, SEP) is sampled at n points, and the number of permitted states at each point is s, the total number of possible trajectories is given by s^n.

commonality (the extent to which sequences represent shared experiences). These attributes may be readily and precisely quantified via the measures introduced in Section 14.6, and discussed above. However, the trajectories of Appendix 5 possess another, more subtle, property which does not lend itself to quantitative representation. This additional concept may be illustrated by examining a number of specific social sequences from those observed in the study. Consider the following group of three trajectories *(interpretation of symbols is as for Appendix 5)*:-

```
      ID                          SEQUENCE
============================================================
      93      00000000.000000111111111111111111111111111111
      94      0000000000000000...111111111111111111111111111
      105     00000000000000000000000000001111111111011111
```

While these trajectories are unique (that is, were each observed for only one subject), they loosely conform to a common general pattern which, importantly, may be concisely expressed: they indicate *upward mobility*, albeit with irregularities, and with differences in when the transition from the manual state took place. A readily-definable common form is also exhibited by the following sequences:-

```
      ID                          SEQUENCE
============================================================
      56      000..000000000000111111111111111111110000000
      70      0000.000000000000000000011111111111111000000
      110     00000000000000000000111111111111100000000000
```

Although these trajectories are also unique, they again conform to a general pattern which lends itself to easy characterisation: they might be described as *peaked*, or as indicative of *temporary upward mobility*. The idea that some of the individualism evident in subjects' social trajectories represents relatively minor deviations from an underlying general pattern has been expressed by other researchers; for example, Wadsworth *et al*. assert:-

> "...diverse aspects of the life course can be characterized in terms of developmental *trajectories* (spanning relatively extended developmental periods), punctuated by *transitions* between individual states. Trajectories may be defined in terms of social statuses such as employment... ...individuals will vary in their relative positions on any specific trajectory. In some instances it may be appropriate to envisage a single underlying 'normative' trajectory, around which individuals deviate." (Wadsworth *et al*. 2007, pp. 9-10)

Now, contrast the two groups of sequences shown above with a further selection of unique trajectories:-

```
      ID                          SEQUENCE
============================================================
      164      00011111100111111100000111111111111111111100
      202      1111..0000011100000000000111100011111111..1111
      210      11111111000000000000000000011111111111100000
```

In contrast to those shown earlier, none of these trajectories represents a specific instance of a general pattern which may be concisely expressed (such as ‘upward mobility’), nor does any of these sequences conform to the most common paths via which individuals might be expected to progress through life. The first set of sequences shown above (93, 94, 105) represents people who (informally stated) ‘go up in the world’, while the second group (56, 70, 110) depicts the experiences of individuals who improve their initially low social position, but revert to more modest circumstances in middle age. There are many plausible explanations for such a pattern. For example, a person may decide to leave a stressful profession, and adopt a more lowly but satisfying occupation as s/he grows older, the transition perhaps being encouraged by a generous voluntary redundancy payment. However, the final group of trajectories (164, 202, 210) is arguably different, in that none of its members appears to match any instantly-recognisable or ‘expected’ pathway through life. It is stressed that the difference between these latter sequences and those in the other sets is not simply one of complexity (as the concept has been defined in this discussion). Rather, the distinction between the final set and the others relates to (a) the recognisability or intuitive plausibility of the experiences involved, and (b) the ease with which the observed patterns may be verbally described. With regard to (b), there are similarities to the concept of Kolmogorov complexity which is central to algorithmic information theory (Li & Vitanyi, 2008). Kolmogorov complexity may be defined as “the minimum number of bits into which a string can be compressed without losing information” (NIST 2009b), or as “the size of the shortest computer program that could generate a complete description of the object”. (Mitchell, 2009). The difference between Kolmogorov complexity and the definition of complexity used in the present discussion may be demonstrated by a simple illustration. Consider two truncated sequences of social position, represented using the now-standard symbolography i.e. 0 = manual, 1 = non-manual:-

```
SEQUENCE A: 00001000010000100001

SEQUENCE B: 00110100011000000111
```

Both sequences contain seven state transitions, so their respective levels of complexity (according to the definition used thus far in this discussion) are equal. However, while

sequence A may be completed represented by a statement such as "repeat the pattern `00001` four times", a similar abstract description of sequence B would be markedly longer. Thus, the Kolmogorov complexity of A is lower than that of B. Extending this to the sample sequences presented earlier, it is clear that an adequate verbal characterisation of any sequence in the final group (164, 202, 210) is more challenging to create, and will be more lengthy than a corresponding description of any trajectory in the other two groups.

The above discussion may be summarised by asserting that the trajectories of SEP observed in this study possess a further (rather ill-defined) property beyond those of complexity and commonality, namely the extent to which they are or are not *unusual* (that is, represent paths through life which might be considered 'surprising' or idiosyncratic). While many patterns of social experience may be regarded as representing minor (and possibly unimportant) departures from intuitively appealing 'archetypal' configurations (such as 'upward mobility'), other trajectories do not conform to any recognised general pattern. This additional property has been discussed at some length, because it is arguably important for two reasons. First, in the specific context of the present study, the existence of such unusual social trajectories lies at the heart of the challenges which arose when investigating associations between SEP and health (see Section 6.3.1). These challenges were essentially twofold: the inability to estimate maximum likelihood parameters (due to sparseness in the data), and the difficulty of interpreting large numbers of parameter estimates. The analytically troublesome diversity of personal experience observed in this study prompted the attempt to reduce that diversity by grouping or clustering the original sequences. However, unusual or anomalous sequences do not by their nature lend themselves to such grouping, with the result that the technical problems encountered during analysis of sparse data (that is, the inability to derive maximum likelihood estimates of regression parameters) cannot be completely resolved by the artifice of clustering. This was illustrated in Section 6.3.7 which explained how, after application of the clustering process to the sequences of social position, several small clusters containing either one or two individuals were retained. The trajectories involved did not fall into any of the larger groups precisely because they were unusual, and had ultimately to be eliminated from analysis.

The problems resulting from the presence of unusual trajectories are important for a second reason, namely that they carry implications for the design of future studies into health inequality. It might initially be thought that the difficulties caused by unusual patterns of experience could be resolved in future research simply by increasing the sample size, which

is an obvious response to problems centred around sparseness in the data. However, a moment's consideration suggests that this expectation is unfounded, because it appears highly likely that increasing the sample size would also increase the probability of observing unusual trajectories. Certainly, the scope for observing unusual trajectories is theoretically vast: with (as here) 46 sampling points and a four-way state space at each point, the number of possible social sequences is given by 4^{46} (= 4.95176e+27); an inconceivably large number. In practice, there is likely to be a ceiling on the number of distinct trajectories observed, in that some of the theoretically possible sequences are very unlikely to be encountered. For example, a trajectory such as

```
01010101010101010101010101010101010101010101
```

would be viewed as very surprising indeed. However, although improbable, such patterns of experience are not impossible, and it is reasonable to hypothesise that one consequence of seeking larger sample sizes would be the detection of a greater number of unusual social trajectories. Thus, the problem of sparseness in the data and the resultant analytical challenge of nonestimability might well be repeated in future studies which collect detailed social trajectories, even if based on samples considerably larger than those available here.

***

At this point, conclusions relating to subjects' social experiences over the lifecourse are summarised.

17.2.5 Lifetime social experiences: summary

The distribution of accumulated lifetime disadvantage which was observed in the study exhibits two main features of interest. First, a marked difference in the respective disadvantage experiences of men and women was indicated. However, interpretation of this difference is problematical. The approach used to calculate disadvantage for women introduced an element of artefactual distortion, the likely effect of which was to make the apparent distribution of disadvantage among females more similar to that for males. Consequently, the (unknown) 'true' degree of difference between sexes in lifetime exposure to disadvantage may in fact be more marked than that indicated in Figure 11.2.1 and Table 11.2.1. Characterising the social position of women (and hence their exposure to disadvantage) in terms of occupational class is fraught with difficulties (see Sections 6.1.1

and 6.1.3), and over-interpretation of the by-sex difference in cumulative disadvantage observed in this study would be unwise. The issue of interpreting sex-related differences in the experience of disadvantage has seldom been confronted in health inequality research, because much work in the area has been limited to men[49]:-

> "Links between low socioeconomic position and poor health are well established. Most previous research, however, has focused on middle-aged males." (Beebe-Dimmer *et al*. 2004, p. 481)

A second feature of interest relating to accumulated disadvantage is that while its distribution among men in this study exhibited a broad resemblance to that observed in two other British studies (Table 17.2.1), the comparison highlighted a potentially important limitation of the classic 'sparse' sampling approach which has traditionally been used to estimate lifetime disadvantage in health inequality research. The discussion in Section 17.2.1 illustrated how a specific state (specifically, that of complete freedom from disadvantage) might be imprecisely captured by a coarse sampling regime which determined SEP at only a small number of widely-spaced time points. As a result, the many studies which have relied on such a representation to assess relationships between SEP and health are susceptible to the criticism that they may have incorrectly estimated the associations of interest. The actual issue of whether absolute freedom from disadvantage is 'important' (that is, whether a health outcome is more or less likely to be observed among those with zero exposure to disadvantage) is largely irrelevant here. What is noteworthy is that the detection of specific patterns of social experience by a study design which has shaped much current knowledge of health inequality may be imperfect.

Further possible limitations of the classic sparse sampling design were illustrated by consideration of the study's findings relating to subjects' explicit trajectories of socioeconomic position. Section 17.2.2 outlined how the average number of social state transitions observed in this study actually exceeded the maximum which can be represented by the classic three-point sampling scheme. In the same vein, Section 17.2.3 demonstrated how the number of unique social trajectories observed greatly exceeded the relatively small number of distinct trajectories which are assumed by sparse sampling regimes. Section 17.2.4 extended discussion of the latter point by suggesting that while many of the

[49] Some researchers have minimised the problems related to classifying women on the basis of OSC by restricting the study sample to women who are currently in employment (e.g. Heslop *et al.*, 2001). This avoids the challenges associated with interpreting the non-employed state for women, but obviously limits the generalisability of the findings to female populations.

trajectories observed in the study actually represent minor departures from certain general patterns (such as upward mobility), some individuals record unusual sequences of SEP which do not conform to any recognised archetype. Under a sparse sampling regime, such a 'perverse' trajectory is effectively forced into identity with the 'standard' trajectory which it most closely resembles.

All of the potential limitations identified above are actually different expressions of a single underlying concern: that the determination of lifetime SEP by infrequent ('sparse') sampling, via which much current knowledge of health inequality has been obtained, may *over-simplify* the reality of social experience over a lifetime. This, as previously stated, raises the possibility that crucial relationships between possibly flawed representations of lifetime SEP and health may have been incorrectly estimated.

***

The issue of the limitations of sparse sampling schemes is returned to later, as part of a wider discussion of the methodological challenges associated with representing sequential categorical data (specifically, trajectories of SEP) in analysis. For the present, discussion proceeds to consider variation in subjects' exposure to residential hazards over adult life.

17.3 Exposure to residential hazards over the lifecourse

17.3.1 Accumulated exposures

In contrast to the broadly U-shaped distribution of disadvantage, the cumulative experiences of dampness (Figure 11.4.1 and Table 11.4.1), air pollution (Figure 11.6.1 / Table 11.6.1) and both combined (total hazard load; Figure 11.8.1 / Table 11.8.1) all exhibit a clear unimodal distribution, with the mode occurring at the zero point (that is, no exposure). For the two individual hazards of dampness and air pollution, the experience of complete freedom from exposure was common. One hundred and sixteen of 220 subjects (= 52.7%) recorded zero exposure to dampness, while the corresponding proportion for air pollution was 118 / 215 (= 54.9%). For the combined measure of total hazard load, 58 of 210 respondents (= 27.6%) received no exposure. In a further contrast with the experience of disadvantage, none of the measures of housing risk differed significantly between the sexes. For all three residential hazards, Mann-Whitney tests of equivalence between sexes returned

nonsignificant values, and the relevant Figures (see above) suggest little variation by sex. The absence of a between-sex difference is entirely plausible, there being no *prima facie* reason why the experience of residential (as distinct from occupational) hazards should not be the same for men and for women.

A noteworthy feature of the residential risk distributions, partly visible in the Figures, is the abruptness of the discontinuity between zero exposure and the adjacent higher (nonzero) values. This effect is demonstrated more clearly in Table 17.3.1, which shows the lower end of each distribution in numerical form. The existence of such clear natural breaks suggests that, for analysis purposes, there may be some justification for treating exposure to these hazards as dichotomous quantities, contrasting the unexposed state with 'any' exposure (i.e. 1 year or more). Had the envelope of the distribution been smoother, with no clear discontinuity at the zero / one point, such a dichotomisation would be harder to justify.

TABLE 17.3.1: Lower end (0 years to 3 years) of distributions of cumulative exposure to residential hazards.

number (%) of subjects reporting....	**dampness (220 subjects)**	**air pollution (215 subjects)**	**total hazard load (210 subjects)**
zero exposure	116 (52.7)	118 (54.9)	58 (27.6)
1 year of exposure	5 (2.3)	4 (1.9)	8 (3.8)
2 years of exposure	9 (4.1)	3 (1.4)	6 (2.9)
3 years of exposure	10 (4.6)	5 (2.3)	11 (5.2)
etc.	...	...	...

Overall, the results obtained in this study suggest that in the cohort represented by this sample, cumulative exposure to dampness and to indoor air pollution over adult working life exhibits three distinctive characteristics. First, the distribution of exposure to each hazard is heavily skewed, with the modal value sited at the zero point and a long right tail. Second, the distributions exhibit marked natural breaks at the 'zero versus any' boundary. Finally, the experience of these hazards is very similar for both sexes.

***

Attention now turns to the trajectories of exposure to residential hazards.

17.3.2 Trajectories of residential hazard exposure

Following the approach used to examine the trajectories of social position, consideration of subjects' time-related exposure to housing hazards begins with the measure of complexity

defined earlier. For exposure to dampness, the observed MTD value (i.e. the average number of state transitions) was 2.04; the corresponding value for air pollution was 1.98. It thus appears that the experience of these exposures was relatively stable; that is, involved few changes in exposure status. These values reflect the high proportions of individuals who experienced consistent freedom from exposure (i.e. no state changes) across the entire period examined (over half of the sample for both hazards; see Section 17.3.1).

In addition to the raw numbers of state changes (which are summarised in the MTD values), the location of transitions in time is also of interest. Some insight into how the prevalence of exposure to both hazards varied with increasing age is provided by Figure 14.3.1 (dampness) and Figure 14.3.2 (air pollution). The former suggests a clear between-sex difference in the experience of dampness, exposure being less common among women at younger ages than among men. This is difficult to account for: there is no *prima facie* reason why females should be less likely than males to report dampness in the home at certain ages. Any explanation of this difference would be purely speculative, and none is attempted here. In contrast, the age-related prevalence of air pollution is strikingly similar between the sexes, a common pattern being evident: general decline from the start of the age range, levelling out to a 'plateau' effect in later middle age.

Interpretation of the findings relating to air pollution is difficult, due to the way in which exposure to this hazard is defined in the dataset (the relevant extract from the *Users Guide to the Dataset* is reproduced in Section 5.3.2). The problem is that the definition is dynamic, in that the nature of the risk changes over time: someone living in proximity to an industrial facility would be considered exposed to pollution prior to 1960, but not exposed thereafter unless s/he lived close to a major road. Thus, a subject's pollution status may change over the 1960 boundary, even though the person continued to live in the same dwelling. The difficulties which this presented in analysis were outlined in Section 7.5. A consequence of this definition is that the pattern evident in Figure 14.3.2 can be considered an accurate representation of subjects' exposure to pollution only if the assertion made in the Users Guide is correct: "Following the Clean Air Acts of the 1950s, the main source of air pollution in the U.K. began to change from industry to road traffic." (*Users Guide to the Dataset*, p.3). In fact, the general decline in pollution exposure suggested by Figure 14.3.2 is plausible, reflecting legislative developments such as the Clean Air Acts of 1956, 1968 and 1993. One study has concluded that the Clean Air Act 1956 did contribute to improving air

quality in the UK over a period of 30 years following its enactment, though other social and structural changes also played a significant role (Giussani, 1994).

Attention now turns to the measures of commonality, which indicate the degree of diversity or homogeneity in individuals' sequential experiences of residential hazards. The observed values for these measures were shown in Table 14.6.1, but are repeated for convenience:-

dampness:	MSF	=	2.24
	SUI	=	0.41
air pollution:	MSF	=	3.03
	SUI	=	0.29

Comparison with the corresponding values reported earlier for social position (MSF = 1.36; SUI = 0.66) indicates that, relative to the latter, subjects' trajectories of exposure to residential hazards exhibited greater commonality. That is, an individual sequence was more likely to be shared among multiple respondents. In fact, such a comparison can only be tentative because there are major conceptual differences between the sequential representation of SEP and those of exposure to residential risk. Notably, the former features a four-way state space at each yearly point (manual, non-manual, Armed Forces and non-employed) while residential risk is represented by a simple exposed *vs.* not exposed contrast. Consequently, it might be argued that the potential for individuality differs between the two types of experience[50]. This is certainly true in an abstract mathematical sense. As shown earlier, the number of possible social trajectories is 4^{46} (= 4.95176e+27), but the corresponding number of residential exposure sequences is 'only' 2^{46} (= 7.036874e+13). Both numbers are of course almost unimaginably large.

To summarise, the trajectories of exposure to specific residential risks which were created for this study exhibit a number of features of interest. First, for both hazards, trajectories were relatively stable, with an average of two state transitions being observed over the 46 year period examined. Second, patterns of exposure to these risks were characterised by less individuality than sequences of social position recorded over the same time period, though the interpretation of this difference is problematical. Third, while the prevalence of exposure to air pollution over time was broadly similar for both sexes, the time-related experience of

[50] Alternatively, it might be argued that the probability of transitioning from one state to another is independent of the number of destination states which are available. Under this view, there is no automatic reason why sequences of SEP (four possible states) should exhibit greater individuality than trajectories of residential hazard exposure (two states – exposed *vs.* not exposed).

dampness exhibited clear differences by sex in young adulthood No *prima facie* explanation for this difference suggests itself.

17.4 Joint experience of multiple hazards

In discussing the observed variation over time in social and residential exposures, each class of hazard has so far been treated in isolation. However, patterns of variation in subjects' sequential exposure to combined hazards are also of interest. A method for visualising respondents' joint exposure to multiple risks (social disadvantage, dampness and pollution) was outlined in Section 10.4, and the results of its application were presented as Figure 14.5.1a (for men) and Figure 14.5.1b (for women). While a number of effects are suggested by these presentations, the most striking is the broad increase with age in freedom from all three exposures (shown in the lowermost striation of each figure). The effect is present for both sexes. The figures emphasise the essentially dynamic nature of the factors which are of central interest in the study i.e. SEP and the experience of housing-related risks. That these should vary over time is entirely predictable: the notion that they might remain static, for all subjects, over a period of four and half decades would not be seriously entertained. This study has been able to identify the nature of time-related variation in these factors to an unusual degree of detail.

***

The discussion of results in this chapter represents an attempt to answer the first research question posed by the study (see Section 4.3). While a number of features of interest have been identified, arguably the most important aspect of these results is the insight which they provide into possible shortcomings of the approaches to representing lifetime SEP which have featured prominently in health inequality research. This topic is now developed further in the context of a discussion focussed on the study's final research question: how may detailed patterns of sequential experience most effectively be represented in health inequality studies? The question is divided for discussion purposes into two sub-themes:-

i. Given the limitations of the sparse representations of SEP which have underpinned much research into health inequality, are more detailed representations (such as

those used in this study) necessarily and automatically superior to the classic approach?

ii. Where detailed sequences of experience are available, how may such data (which effectively consist of a categorical time series) be represented in statistical analysis?

Both sub-questions are discussed in the next chapter.

CHAPTER 18: DISCUSSION (III) – REPRESENTING SEQUENTIAL PATTERNS OF RISK EXPOSURE IN LIFECOURSE ANALYSIS

18.1 Introduction

A fundamental postulate of the lifecourse approach to explaining health inequality is that exposures experienced across an extended period of time exert an influence on health. This is true irrespective of which of the two main 'strands' of the lifecourse approach (critical period, or accumulation of risk; see Section 2.2) is held to apply. The central idea has been expressed thus by two of the foremost authorities in the field:-

> "We have defined a life course approach to chronic disease epidemiology as the study of long-term effects on chronic disease risk of physical and social exposures during gestation, childhood, adolescence, young adulthood and later adult life." (Ben-Shlomo & Kuh 2002, p. 285)

A key requirement of investigations which postulate lifecourse influences on health is that the putative risk factors (the 'physical and social exposures' of Ben-Shlomo & Kuh) are represented adequately; that is, that the exposures of interest are collected, to an acceptable degree of precision, across all or most of subjects' lives. This requirement has been explicitly identified by Davey Smith *et al.*:-

> "Any serious attempt to elucidate the contributions of socially distributed risk factors to the risk of disease in adulthood should aim to collect information covering the entire lifespan of study participants." (Davey Smith *et al*. 1997, p. 552)

Discussion thus far of the present study's findings has argued that lifetime social exposures (specifically, the experience of disadvantage) have often been measured in research by sampling schemes which may not fully capture the individuality of lived experience over a period of several decades. The present chapter extends this discussion by considering in more detail the respective merits of the classic sparse sampling regime which has been used extensively in inequality research, and more frequent (or 'dense') sampling, as used in the present study. This comparison is then developed further into a methodological discussion of how sequential experiences may most effectively be represented in epidemiological analysis. Throughout, the discussion focuses on *social* exposures, but much of the argument is equally applicable to the experience of environmental risks (such as the housing hazards considered in this study).

18.2 Representing lifetime experiences: sparse sampling *vs.* dense sampling

Section 17.2 summarised the potential limitations associated with the infrequent sampling of lifetime SEP by postulating that sampling schemes of this type over-simplify the richness of actual experience. Specifically, such schemes may

- be unable correctly to detect specific patterns which may be of particular interest (e.g. the experience of complete freedom from social disadvantage)
- fail to represent the number of social state transitions (e.g. *from* manual *to* non-employed) which are experienced in the lives of some individuals
- not accommodate the full diversity of social trajectories which exists

These potential limitations were illustrated with reference to the results of the present study, which captured SEP over adult life to a considerably greater degree of temporal precision than is commonly the case. The implied conclusion is that frequent or dense sampling of SEP of the kind used here is superior to the classic sparse sampling which has been the workhorse of inequality research. Dense sampling undeniably collects more raw information, thus in theory permitting associations between lifetime social exposures and health in later life to be assessed more precisely. However, it may be argued that the apparent superiority of frequently collecting SEP data is actually questionable, for two reasons which are now considered in turn.

The first of these is simple, and relates entirely to the practicalities of assembling data which represent a person's socioeconomic position (and include other relevant covariates) at frequent intervals over life (e.g. yearly). Collecting such data prospectively on an *ad hoc* basis is obviously not feasible: decades would elapse before the required information was compiled. One alternative is the retrospective elicitation of lifetime data (as used in the present study); another is reliance on information progressively accumulated by existing long-running cohort studies. Each of these has its drawbacks. Reliance on existing cohort studies is entirely dependent on availability of the precise data elements required for a specific investigation, which will often not be the case. The retrospective elicitation of information assumes the accuracy of subjects' memories, and their willingness to fully disclose their recollections. Furthermore, individuals recruited into such studies may be nonrepresentative of the populations of interest in important ways. To illustrate, it has been acknowledged that the dataset used in the present study consists of people who were possibly

less disadvantaged, and enjoyed better health, than the notional population from which they were drawn. Despite such limitations, retrospective information on lifetime SEP has been successfully used in major epidemiological studies. An example is provided by the SHEEP study (Stockholm Heart Epidemiology Program) (Reuterwall *et al.*, 1999; Hallqvist *et al.*, 2004; Ljung & Hallqvist, 2006). This collected, by questionnaire, retrospective information on subjects' occupations (including job titles and details of tasks performed) throughout working life.

One further approach to obtaining detailed information on lifetime social exposures is to use routine government data relating to citizens, but this is possible only in jurisdictions where relevant data are both systematically collected and readily made available to researchers. One example is provided by a study of Naess *et al*. (2004b), which examined links between SEP and cause-specific mortality in Norwegian men. In this study, SEP in childhood was determined by housing conditions (as recorded in the 1960 Census), while adult socioeconomic position was assessed on the basis of income information provided by the taxation authorities[51].

Relatively straightforward practical concerns (such as those discussed above) are not the only reason for querying the superiority of frequent sampling of SEP over the traditional (infrequent) approach. A second argument against dense sampling, applicable only when used to construct trajectories (as distinct from cumulative measures), is that the information so derived is difficult to exploit in many kinds of statistical analysis without being subject to some form of condensation or reduction. A measure of accumulated disadvantage constructed (as in this study) from 46 yearly points is a simple quasi-continuous (integer) quantity, and presents few problems in analysis. Such a measure is almost unquestionably a more accurate representation of the person's disadvantage experience than one based on (say) three widely-spaced points, and is thus of undisputed value. However, a *trajectory* built from 46 points is effectively a categorical time-series which, while providing a more precise portrayal of the subject's 'true' pattern of experience over time than one constructed from three or four points, presents formidable analytical challenges.

Section 6.3.1 in the *Methods* section discussed the analytical problems presented by the social trajectory data, specifically nonestimability of parameters via maximum likelihood

[51] Norway provides ideal conditions for such studies, because completion of the national Census is compulsory, and information from various governmental sources may be reliably linked via the Norwegian identity code.

(due to sparseness in the data), and the difficulty of interpreting large numbers of individual effects. Of these, issues relating to sparseness are almost unavoidable when dealing with appreciable numbers of distinct trajectories. To illustrate, the previously-cited study by Naess *et al*. (2006) collected a binary representation of SEP at four points, yielding 16 distinct trajectories. This study included 7,891 male individuals; however, four of the 16 trajectories were observed for, respectively, 58 subjects (i.e. 0.7% of the total), 72 subjects (= 0.9%), 86 subjects (= 1.1%) and 88 subjects (= 1.1%). Thus, even in a large study with a small number of trajectories, the problem of sparseness begins to be discernible. In the present study, this problem was massively amplified: 294 subjects yielded 216 distinct trajectories, and the data were clearly unusable in their original form.

The method used to condense the original trajectory structure into a more analytically tractable number of groups was described at length in Section 6.3. However, the results of this process were to some extent unsatisfactory. Examination of the final social cluster structure (Appendix 2) reveals some definite anomalies. For example, Cluster 3 broadly consists of subjects who were mainly in the manual state throughout, but includes the sequence `0111000000` which incorporates a contiguous block of 15 years spent in the non-manual condition. Similarly, it is difficult to provide any concise characterisation of Cluster 7, which partly consists of 'peaked' trajectories (i.e. manual ➔ non-manual ➔ manual), but includes sequences which do not conform to this pattern. One interpretation of such features is that despite being based on data which are far more detailed than those commonly available in health inequality studies, the clustered scheme of Appendix 2 suffers from essentially the same limitation as previous research: it *over-simplifies* (and thus distorts) individuals' social experiences.

However, the presence of such anomalies does not mean that the actual method used in this study to group the original social trajectories (i.e. clustering based on distances derived via optimal matching) is irretrievably flawed. The same method was applied to the other sequential data used in the study (that is, the sequences of exposure to dampness and to air pollution) with some success. For example, the air pollution groups identified in Appendix 4 are readily interpretable, each cluster exhibiting a definite homogeneity of experience in its constituent sequences. The groups might loosely be interpreted as follows:-

Cluster 1 - complete freedom from exposure (to air pollution)
Cluster 2 - minimal exposure

Cluster 3	-	exposure in early adult life, succeeded by complete freedom from exposure
Cluster 4	-	sustained exposure for first half of period examined; freedom from exposure thereafter
Cluster 5	-	consistent or near-consistent exposure
Cluster 6	-	no exposure in first half of period; sustained exposure thereafter
Cluster 7	-	freedom from exposure in first 10 to 15 years of period examined; consistent or near-consistent exposure thereafter
Cluster 8	-	exposure for all but the last 10 to 15 years of the period examined

Although the specific approach used to condense the sequences met with varying degrees of success, the important point is that some form of reduction or condensation of the original trajectories was essential to permit analysis. Other possible approaches to condensing such data are considered later in this chapter, but the recognition that such condensation (however achieved) is necessary leads to an interesting conclusion. Although limitations of the sparse sampling schemes traditionally used in health inequality research have been demonstrated, it might be argued that dense sampling (as used in the present study) is actually no better in practical terms, because data collected by the latter approach are unusable without some form of reduction or condensation. On this basis, a possible argument *against* dense sampling is actually suggested, namely that it may be inefficient and wasteful because the information collected is of little practical value in its original form. In other words, a law of diminishing returns may apply: the substantially increased cost (in time, effort and resources) of collecting dense data may yield little increase in *usable* information relative to that which could be obtained by traditional sparse sampling. A very loose analogy may be drawn with the collection of biomedical measures in research. A quantity such as blood pressure or forced expiratory volume may be assessed by taking the average of two or three readings (in the interests of minimizing measurement errors, or to allow the subject to become comfortable with the measurement regime), but would there be any merit in taking an average of (say) 30 readings? Transferring this question to the context of the present study, is sampling a social trajectory at yearly intervals across several decades really more useful than capturing a few 'snapshots' at widely-spaced intervals?

The conceptual *impasse* identified above (i.e. the conflicting merits of sparse and dense sampling) is not pursued further at this stage, but is returned to in Chapter 19. Discussion now turns to a methodological issue suggested by the conclusions expounded thus far. It has been demonstrated that where (as in this study) detailed sequential data representing life-

time exposures are available, some form of condensation or reduction will almost certainly be required to permit analysis of the observed trajectory structure. The specific method used in this study yielded mixed results, raising the obvious question of whether alternative methods exist via which the desired goal of condensation might be achieved. Two possibilities are discussed in the next section. Interest here is not restricted to methods which might realistically have been used in the present study. Rather, the intention is to consider how other analytical approaches could be applied to highly detailed sequential data (representing either social or environmental exposures) which might be available to future research in lifecourse epidemiology.

18.3 Alternative representations of sequential data in analysis

18.3.1 Semiparametric group-based methodology

The solution adopted in this study to cater for the problem of 'excessive diversity' had as its goal reduction of the *n* individual trajectories identified in the data into a smaller number of groups or clusters. The aim was to derive a cluster structure characterised by high degrees of within-group homogeneity, and of between-group heterogeneity. That is, trajectories within a group would be similar to each other, but different from trajectories assigned to other groups. Conceptually, this structure was envisaged as representing a typology of experience over time, the typology ideally being sufficiently compact as to permit effective analysis. One alternative approach to achieving the same goal involves the use of semiparametric group-based methodology or SPGM (Land & Nagin, 1996; Nagin, 1999; Nagin & Tremblay, 1999; Nagin, 2005). SPGM is in essence a technique "designed to identify relatively homogeneous clusters of developmental trajectories." (Nagin 1999, pp. 139-140). That is, it identifies groups of individual trajectories (of experience or, more typically, of behaviour) which share common characteristics. The group scheme derived via SPGM may then be used as a predictive factor in regression-type analyses. For example, Nagin & Tremblay (1999) used SPGM to identify behavioural trajectories among boys, the grouped trajectories then being used in regression models to predict juvenile delinquency.

Consideration of this study by Nagin & Tremblay helps to illustrate the potential application of the method in lifecourse investigations of health inequality[52]. These authors identified, via the application of SPGM, four distinct trajectories of problem behaviour. The first (labelled 'lows') consisted of individuals who "rarely display the problem behaviour to any substantial degree". In the context of an inequality study, these might correspond to people with minimal exposure to disadvantage. A second group ('moderate-level desisters') manifested modest levels of the behaviour at an early age, but largely desisted as they grew older. These are analogous to upwardly mobile individuals, who experience some disadvantage at the start of the lifecourse, but progress to the non-disadvantaged state at a fairly early point in life (say, young adulthood). A third trajectory group ('high-level desisters') exhibited high levels of the behaviour when young, but progressed to far lower levels later. However, the 'desisting' phase was located later in life than was the case for moderate-level desisters. Thus, the high-level desisters may be compared to people who are subject to extreme disadvantage in early life, but largely emerge from this state in (say) middle age. A final group ('chronics'), who "start off scoring high on the behavior and continue to score high throughout the observation period" form an obvious parallel with those who suffer constant disadvantage throughout life.

The SPGM framework has been extensively used in criminology and psychology, but a recent paper by Sturgis & Sullivan (2008) is of direct interest to the present study, in that it reports an application of the method to identify social class trajectories. Using data from three waves of the 1970 British Cohort Study (BCS70), these authors applied SPGM to identify five qualitatively distinct trajectories of occupational class among male subjects[53]. This work, which was not available when the present study was embarked on, offers a potential alternative to the approach used in the study to increase the analytical tractability of the sequence data shown in Appendices 5, 6 and 7.

One restriction of SPGM is that models within this framework are restricted to continuous or ordinal outcomes. Consequently, the method could not be used directly with the representations of SEP and residential hazard exposure which were developed for the present study, these being based respectively on a four-way nominal scheme (manual, non-manual,

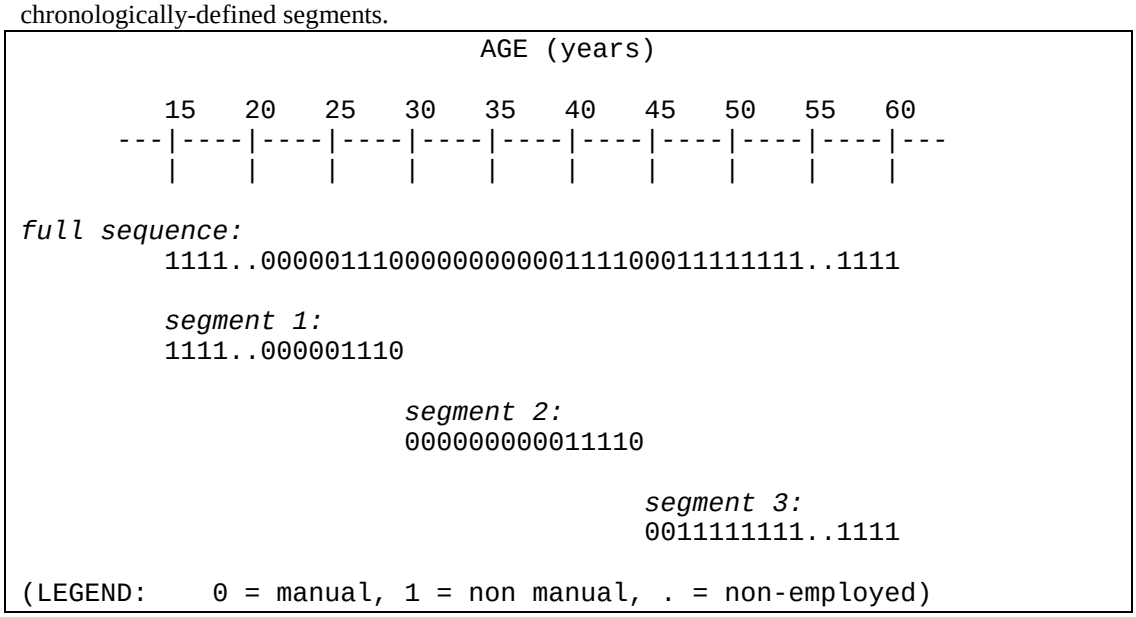
[52] All quotations in this paragraph are from Nagin & Tremblay (1999), p. 1189. To ease reading, the normal convention of citing the source of each quotation is dispensed with in the paragraph.
[53] Throughout this paper, the authors actually use the term 'latent class growth analysis' (LCGA) when discussing their method, but also recognise the alternative designation: "LCGA... ...is also referred to as 'semiparametric group-based trajectory analysis'". (Sturgis & Sullivan 2008, p. 70)

Armed Forces and non-employed) and a simple binary contrast (exposed *vs.* not exposed). There are a number of ways in which the sequential representations of these factors could be modified to facilitate the use of SPGM. In the case of the social location data, one possibility would be to abandon the reduction of the occupational class schema into a simple manual / non-manual dichotomy, instead retaining the original classes (I, II, IIINM etc.) which may be treated as an ordinal quantity. This was in fact the approach adopted by Sturgis and Sullivan. However, problems would remain with the Armed Forces and non-employed states, which cannot be accommodated within an ordered scheme. In any event, this method would not help with the application of SPGM to the residential hazard data, which consist of binary indicators at each age point.

An alternative way of reshaping the data for use with SPGM is now given in brief outline, the method being applicable to both the social and residential factors featured in this study. The approach involves two stages. First, each sequence of year-on-year experience is divided into *n* segments, all segments being of approximately equal length. A simple illustration of the process is given in Figure 18.3.1, which shows a sequence of social position being divided into $n = 3$ segments. The first of these contains the ages from 15 to 29 (15 yearly points), while the remaining segments consist of, respectively, ages from 30 to 44 (15 points) and from 45 to 60 (16 points). The second stage of reshaping involves identifying the proportion of the total number of years in each segment which is represented by what may be termed the *state of interest*. For the social position data, the state of interest is considered to be social disadvantage (defined as any of manual, Armed Forces or non-employed[54]), while for the two residential hazards the state of interest is the exposed condition. In segment 1, the proportion of yearly points indicating the disadvantaged state is $8 / 15 = 0.53$, while the corresponding proportions for the remaining two segments are given by $11 / 15 = 0.73$ (segment 2) and $4 / 16 = 0.25$ (segment 3). In this way, the original sequence of 46 nominal elements is reshaped into a quasi-continuous quantity representing exposure to social or residential risk which is effectively sampled at *n* different time points. Data in this form are suitable for analysis via SPGM. As with the method actually used in the study, information is (intentionally) lost in a trade-off to achieve greater usability of the data in analysis.

[54] See Section 6.2 in the *Methods* area.

FIGURE 18.3.1: Illustration of dividing a full-detail sequence of social location into three chronologically-defined segments.



18.3.2 Functional data analysis

Functional data analysis (Ramsay & Silverman, 2005; Ferraty & Vieu, 2006) is "a recently developed method of analysing data consisting of serial measurements, where each data series, or growth curve... ...is termed functional data." (Cole 2007, p. 151). The principle behind functional data analysis (FDA) is that data which are originally represented as a series of discrete values are converted into a smooth curve which is treated for analytical purpose as a single entity. In the context of a study such as the present one, this would correspond to treating each individual's trajectory of experience over time as a single datum or observation. Functional data may serve as either the dependent or independent variable in regression-type analyses (Cole, 2007). One advantage of the method is that the number of measurements, and the locations in time at which they are captured, may vary between analytical units. Thus, trajectories of SEP would not need to be sampled at regular intervals (such as yearly). Ramsay & Silverman (2005) have illustrated functional data drawn from a wide range of research areas, including:-

- the heights of girls measured at ages between birth and 18 years
- mean monthly temperatures observed at a number of Canadian weather stations
- angles in the sagittal plane formed by the hip and knee of children going through a gait cycle

- biomechanical data representing the force exerted on a meter during multiple repetitions of a 'pinch' manoeuvre executed by the human thumb and forefinger

The method of FDA could not be applied to the data used in the present study, because the latter do not consist of repeated measurements of a continuous quantity. However, a study which expressed SEP as (say) income, or which characterised exposure to residential risks by assessing the severity of dampness on a numerical scale with a reasonable degree of resolution (e.g. from zero to ten, or via a visual analogue scale), would in theory permit analysis via the method of FDA. However, a number of potential challenges exist. For example, functional data incorporate the assumption that abrupt variations over time in the quantity of interest are uncommon:-

> "...we usually want to declare that the underlying function is smooth, so that a pair of adjacent data values y_j and y_{j+1} are necessarily linked together to some extent and unlikely to be too different from each other. If this smoothness property did not apply, there would be nothing much to be gained by treating the data as functional rather than just multivariate." (Ramsay & Silverman 2005, p. 38)

It is easy to see how this assumption of limited change between adjacent sampling points could be violated in a study such as the present. For example, if SEP were measured in terms of income, the change in the subject's income due to the onset of unemployment (or the reverse: the acquisition of well-remunerated work after a period of unemployment) could be abrupt. A related problem is that of ensuring that the sampling rate or resolution of the raw data is adequate. Where the value of the function is changing rapidly (as in the instances just cited), the FDA approach requires enough data points to estimate the function effectively. This implies that sampling income at fixed intervals of (say) five years might not be adequate: rather, some form of 'adaptive' sampling would be preferable, in which the sampling rate was increased during periods when the factor of interest was changing rapidly. This would be extremely difficult to implement in practice. Yet another problem is that, under certain circumstances, the estimated function can evaluate to impossible values. For example, if a residential risk such as dampness were measured via a scale from zero to ten, the function might return negative values at certain points, which would clearly be meaningless. Ramsay and Silverman provide a specific example of this effect, in the form of estimated negative rainfall observed by functional smoothing of precipitation data recorded at a weather station.

Despite these difficulties, FDA offers another possible approach to managing serial measurements of exposure in lifecourse investigations of health inequalities. The potential of the method is recognised by Cole in an article discussing relationships between growth curves and health outcomes in later life: "Functional data analysis provides a powerful tool for extending the life course plot... ...This is an area with considerable promise." (Cole 2007, p. 153)

***

In addition to those introduced above, other approaches to exploiting detailed sequential data representing exposure to social or environmental hazards suggest themselves. While these are not considered here, one alternative which is particularly applicable to serial measurements of SEP merits a brief mention. In contrast to the data-driven approach to classification which was used in this study, it would be possible to define a set of classes or categories *a priori*, and effectively coerce each actual (observed) trajectory into the pre-defined class which it most closely resembles. Under this approach, archetypal patterns based on those observed in previous studies would be presumed to exist. For example, Adams *et al*. (2004) identified four social trajectory types (stable non-manual, upward, downward and stable manual), while the eight possible trajectories derived from a three-point sampling scheme have been defined by Hallqvist *et al*. (2004). Forcing a set of observed trajectories into such a scheme would involve the use of classification in the alternative sense introduced in Section 6.3.2:-

> "Classification pertains to a *known* number of groups, and the operational objective is to assign new observations to one of these groups." (Johnson & Wichern 1992, p. 573)

***

At this point, detailed discussion of the study's findings in respect of the three research questions defined in Section 4.3 concludes. Chapter 16 summarised the study's findings in respect of the second research question, considering the overall validity of the conceptual model proposed in Section 4.2 (i.e. that housing conditions mediate the relationship between SEP over adult life and health). The first research question, concerning the nature of time-related variation in exposure to social and residential hazards, was treated in Chapter 17.

Finally, Chapter 18 covered the remaining research question, discussing how data representing sequential patterns of risk exposure may be used in epidemiological analysis.

The next chapter (the last in this thesis) attempts a final synthesis of these findings, summarising the main conclusions of the study.

CHAPTER 19: CONCLUSIONS

19.1 Introduction: presentation of conclusions

Final conclusions drawn from the study's findings are presented separately for each of the three research questions investigated (see Section 4.3, Box 4.3.1). Consideration of each question terminates in a definitive statement in the format common in academic journals: a concise indication of (i) what is already known on the topic, and (ii) what this study adds. The treatment of questions adheres to the same ordering as was used in the *Discussion* chapters 16 to 18 (i.e. Question 2 first, then Q1, followed by Q3).

19.2 Socioeconomic position, the residential environment, and health (Research Question 2)

As stated earlier, the initial motivation for this study was the desire to test the hypothesis that socioeconomic position over adult life may influence health via a mediating effect of housing conditions. Expressed in very informal terms, the hypothesis is that poorer people live in inferior housing; that such housing exposes them to higher levels of specific hazards; and that this elevated exposure results in poorer health. The discussion of results in Chapter 16 demonstrated that this study found no evidence supportive of this view. Possible reasons for the failure to detect convincing associations have been considered in some depth; these are not rehearsed in detail here, but some more general discussion of the extent to which the largely negative findings may reflect limitations of the study is now offered.

The implications of the small sample size available in the Boyd Orr dataset have been acknowledged earlier in this thesis; the relevant passage from Section 5.2 merits reproduction here:-

> "In adopting the dataset for use, analytical limitations - specifically, a lack of statistical power - imposed by the small available sample size (294 cases) were recognised and accepted as unavoidable."

When considering the study's findings, is appropriate to re-emphasise the possibility that the overwhelmingly negative results may be partly attributable to a lack of available statistical power, manifested as Type II ('false negative') errors. Limited (though highly informal) support for this view is given by the observation that many of the point estimates of effects,

although not significant at the conventional 5% level, are consistent with the conceptual model postulated in the study. For example, in Table 13.5.1 (which shows the estimated effect of accumulated exposure to housing risks on the presence of cardio-respiratory disease), ten of the twelve odds ratio estimates exceed the value one; that is, are consistent with the hypothesis that cumulative exposure to poor housing is associated with elevated odds of disease. Scrutiny of the confidence intervals around these estimates indicates that (with one exception) they cannot be interpreted as statistically meaningful results. However, the marked consistency of direction might be viewed, albeit with considerable caution, as a very loose and informal indication that sample size limitations (and the concomitant lack of statistical power) may have contributed to the almost complete absence of associations observed in the study.

Sample size limitations are of especial concern in connection with those specific investigations which involved trajectories of experience (as distinct from cumulative exposures). Acknowledgement that trajectory-based approaches require sizeable samples ('large data') is expressed in the quotation from Naess *et al*. (2006) which was presented in Section 17.1. The sample size implications of permitting greater numbers of trajectories are confirmed by Hallqvist and colleagues:-

> "Could the problem [of distinguishing between critical period, accumulation of risk and social mobility effects] be solved by more information - that is, separating the basic model into more socioeconomic positions at many more periods during the life course? It would rapidly increase the number of trajectories *demanding large numbers of observations*." [emphasis added] (Hallqvist *et al*. 2004, p. 1560)

Ultimately, the influence of lack of statistical power on the results of the study can only be conjectural. However, it is accepted that the sample size offered by the Boyd Orr dataset is arguably too small to permit detection of all but the most gross effects, particularly for those analyses which required that subjects be effectively 'stratified' into trajectories (or groups of trajectories).

The limited statistical power just discussed represents only one of a number of weaknesses in the Boyd Orr dataset. Other limitations of this resource include the atypical composition of the sample which was highlighted in Section 6.2.2, and the highly subjective (and arguably imprecise) measures used to quantify subjects' self-reported exposure to the residential hazards of interest (see Section 5.3.2). Overall it might reasonably be concluded, with the benefit of hindsight, that the dataset was to some extent inadequate for the use to which it

was put in this study. However, this may be countered by arguing that investigation of the conceptual model postulated in the study required a dataset with very rare and specific characteristics: for all its limitations, the Boyd Orr dataset was the best (or 'least bad') resource available.

Beyond limitations of the data source itself, it is acknowledged that the observed lack of associations may reflect methodological weaknesses in the conduct of the study. As in much epidemiologic research, the possibility that confounding influences may have biased the findings cannot be discounted. However, it is argued that because the study considered to some extent the possible influences of two of the most obvious potential confounders - occupational hazards and smoking - confounding is unlikely to have exerted a major influence on the results. In any event, the most likely effect of unmeasured confounding factors would be further dilution of the already extremely weak effects which were observed. A more serious methodological limitation revolves around the rather crude approach used in the study to represent the concept of *risk*. In assessing relationships between social disadvantage and health, and those between housing conditions and health, no allowance was made for the possible influence of an extended induction period (that is, "the period between causal action and disease initiation" [Rothman 1981, p. 253]). It is conceivable that exposure to (for example) damp for a period not exceeding some critical threshold (say X years) may not result in disease, but an exposure of Y years ($Y > X$) would lead to an adverse health outcome, with or without some form of dose-response relationship, once the threshold was passed. The study did not cater for the possibility of such complex non-linear or 'step' effects in the exposure / outcome relationship. In the same vein, the study as executed was not able to allow for critical period effects (that is, variations in vulnerability to a specific exposure at different points in the lifecourse[55]).

Overall, the possibility must be acknowledged that the study's findings relating to the postulated conceptual model may reflect limitations of the dataset, of the methods adopted, or both. Irrespective of the extent to which such limitations apply, the actual findings relating to the model cannot be viewed as other than decisively negative. Accordingly, the relevant conclusions may be summarised as shown in Box 19.2.1 *(next page)*.

[55] In defence of the study, it must be stated that the original conception did include consideration of critical period effects (see the discussion of the timing, duration and fragmentation of exposure in Section 4.1). However, limitations of the dataset prevented this intention from being realised.

BOX 19.2.1: Research Question 2 - statement of current knowledge, and what this study adds.

what is already known on this topic:

Exposure to adverse residential conditions is associated with unfavourable health outcomes, in respiratory health and (to a lesser degree) cardiovascular health. On this basis, a mediating influence of housing conditions has been proposed as an explanation for the persistent phenomenon of social inequality in health.

what this study adds:

Investigation of individual associations linking SEP, housing conditions and cardio-respiratory health identified no consistent framework of relationships among these factors. Consequently, the hypothesis that the residential environment mediates the SEP / health relationship is not supported in the data used for this study.

19.3 Social and residential experiences over time (Research Question 1)

Assessment of the associations involved in the model extended to an investigation of the study's first research question: how both socioeconomic position (hypothesised to be a determinant of health) and exposure to residential hazards (a postulated mediating factor) vary over adult life. Subjects' social and residential exposures, expressed both as cumulative measures and as trajectories of experience, were determined in detail (to the level of the individual year), and their behaviour over time examined.

Discussion of the relevant findings was presented in Chapter 17, and is not repeated in detail here. However, two features are considered sufficiently noteworthy as to merit highlighting at this concluding stage of the thesis. The first of these relates to the observed experience of lifetime social disadvantage. The distribution of this quantity exhibited a distinctly bimodal structure, which carries interesting implications for the nature of the relationship between SEP and health. It is widely accepted that this relationship manifests itself as a *gradient*; that is, health declines steadily and monotonically (but not necessarily in a linear fashion) with lower SEP. This phenomenon was discussed in Section 1.1, and has been identified by, *inter alios*, Marmot *et al.* (1997a), Davey Smith *et al.* (1994) and Bartley (2004). A quotation from the latter describing the effect was presented in Section 1.1, and Davey Smith *et al.* have asserted that "The 'fine grain' of socio-economic differentials in health parallels the fine grading of the social structure." (Davey Smith *et al.* 1994, p. 140). Now, under the widely-accepted view that SEP actually *influences* health (whether directly or indirectly), it

may reasonably be expected that in some cases the distribution of health would reflect the bimodal distribution of lifetime disadvantage suggested by this study, exhibiting the degenerate U-shaped structure evident in Figure 11.2.1. Seeking evidence of such a structure in a range of health outcomes would constitute a substantial research effort, and was not considered as part of this study. Nevertheless, this might be a justifiable research objective for a future study.

A second noteworthy finding revolves around the insights which the study yielded into a class of study design which has been extensively used in inequality research, namely that of sampling the subject's SEP at a small number of widely-spaced points over the lifecourse. The resultant 'sparse' data are then used to create an estimate of accumulated lifetime exposure to disadvantage or (more rarely) an explicit system of social trajectories. This design has historically been of considerable importance in health inequality research: much of the current knowledge in the field has been derived from studies of this type. By creating detailed year-on-year representations of subjects' SEP over adult life, the present study permitted an evaluation of how accurately such sparse sampling schemes may capture an individual's true social pathway through life. The results discussed in Chapter 17 indicate that such schemes simplify lifetime social experiences, to the extent that potentially important features (such as complete freedom from social disadvantage) may go undetected. However, despite demonstrating limitations of this classic approach to representing SEP in inequality research, the study's findings do not lead to the conclusion that frequent or 'dense' sampling of socioeconomic position offers a superior alternative. This is treated further in the next section; conclusions relevant to the study's first research question are summarised in Box 19.3.1 *(next page)*.

BOX 19.3.1: Research Question 1 - statement of current knowledge, and what this study adds.

what is already known on this topic:

(i) The experience of social disadvantage exhibits a graded relationship with many dimensions of health: progressively lower SEP is linked to poorer health status.

(ii) Much current knowledge in the field of health inequality has been derived from studies which sample the individual's SEP at a small number of time points, and use this sparse information to estimate the lifetime experience of disadvantage. The implicit assumption is that such sampling schemes represent a person's true lifetime social experiences with acceptable accuracy.

what this study adds:

(i) The distribution of lifetime social disadvantage may display a bimodal pattern, representing a degenerate 'U' shape. Under the assumption that SEP exerts a causal influence on health, evidence of such a structure in a number of health outcomes would be expected. A systematic search for such evidence represents a possible future research opportunity.

(ii) Detailed representations of SEP over adult life which were created for the study suggest that the widely-used 'sparse' sampling approach may result in over-simplification of subjects' true social experiences, to the extent that potentially important patterns of experience may be undetected.

19.4 Representing sequential experiences in lifecourse analysis (Research Question 3)

Chapter 18 discussed results relevant to the study's third research question. In doing so, two sub-themes were identified. The first of these concerned whether, in view of the limitations of the sparse representations of SEP which have underpinned much research into health inequality, more detailed measures such as those used in this study should automatically be considered preferable. The second theme, motivated by the methodological issues discussed in Section 6.3, centred on alternative ways of representing categorical time series data (specifically, social and residential trajectories) in statistical analysis.

Discussion of the first of these themes (Section 18.2) suggested that, despite apparent limitations of the classic sampling approach, highly detailed sequential data such as those used in the present study do not necessarily and automatically offer a superior alternative. Apart from the practical difficulties of sourcing such detailed data, their exploitation in analysis presents substantial challenges. These challenges were illustrated in this study by introducing the concepts of commonality (the extent to which trajectories are shared among

individuals) and complexity (the degree of flux or turbulence exhibited by trajectories). A third property, introduced in Section 17.2.4, characterised some trajectories as idiosyncractic, unusual or 'perverse'. This latter property was identified as central to the methodological challenges of analysing sequential data, in that it hinders efforts to increase the tractability of such data by classification or clustering approaches such as that used here. The experience of this study suggests that reducing a diverse set of detailed sequential experiences to a manageable number of higher-level groups may be challenging, because some patterns of experience will rarely be encountered, and will therefore be resistant to grouping. As a result, attempts to group trajectory data may well involve compromises (such as deleting unusual trajectories), leading to precisely the same problem of over-simplification which was identified as the real limitation of the sparse approach.

Overall, the study suggests that despite their limitations, the classic sparse sampling regimes which have provided much current knowledge of the health inequality phenomenon could not usefully be supplanted by more sophisticated and demanding approaches such as that used here. The cost and effort associated with the collection of detailed data, coupled with the formidable difficulties of using such data in analysis, mean that approaches of the kind used in this study are unlikely to see widespread adoption. This being so, the present study can perhaps be viewed partly as a methodological experiment. While not generating substantive results of any great significance, the study tested the well-established sparse sampling model against a more complex alternative, and concluded that the former is broadly vindicated. There may be specific investigations for which the application of frequent sampling (whether of SEP or of environmental exposures) is warranted, depending on the research questions of interest. However, studies using this approach are likely to remain rare.

One further feature of the traditional infrequent sampling paradigm, not previously discussed, is that (crudely speaking) it *works*. Studies using schemes of this kind (including the examples cited in Section 17.1) have successfully identified relationships between SEP and various dimensions of health. Indeed, as previously stated, much of the corpus of current knowledge in the area of health inequality has been yielded by such studies. However, this argument cannot be pressed too strongly, for it embodies a logical weakness; specifically, a kind of circularity. Stating that infrequent sampling of SEP works because it predicts health assumes that the associations involved are being estimated correctly, which will be the case only if the representation of SEP used (which is based on infrequent

sampling) is accurate. However, the *persistent* success of infrequent sampling does add weight to arguments in favour of its usefulness and validity.

The apparent adequacy of sparse sampling of SEP raises the further question of *why* such schemes appear to work. The present study suggests that trajectories of lifetime socioeconomic position are both highly individual and complex; consequently, it might be expected that the coarse sampling regimes which have often been used cannot realistically represent a person's social experience. One potential explanation for the apparent success of the sparse sampling approach revolves around the possibility that much of the individuality and complexity in social trajectories represents relatively unimportant 'noise' superimposed on the underlying 'signal' of a common or normative trajectory (such as upwardly mobile). This idea is expressed in the quotation from Wadsworth *et al*. (2007) which was given in Section 17.2.4. If this view is accepted, it may be postulated that the large sample sizes used in many studies which feature sparse sampling effectively negate or 'smooth' some of the noise, isolating the underlying structure. That is, although individual subjects may exhibit minor deviations from a normative trajectory at different age points, these deviations will tend to 'cancel each other out' when subjects are considered *en masse*. This possibility is not considered further in this thesis, but might be explored using a simulation approach. The detailed sequences of SEP created for this study (Appendix 5) could be sampled at (say) three time points – thus effectively simulating a sparse sampling scheme – and the sensitivity of the resulting trajectories to the precise location in time of the sampling points assessed. This would provide some insight into how accurately infrequent sampling can identify the trajectory structure which is revealed when detailed year-on-year information is actually available.

The possibility that, under certain rare circumstances, it might be appropriate to capture the lifetime experience of SEP or some environmental exposure to a fine degree of temporal detail leads to the second sub-theme covered in Chapter 18. Recognising the problems associated with a classification-based approach to representing sequential data, two alternative methods of relatively recent provenance (semiparametric group-based methodology and functional data analysis) offer possible alternatives. However, each imposes specific limitations on the type of data with which it may be used. In general, there is little doubt that the continuing development of lifecourse epidemiology will stimulate ongoing methodological innovation. A good recent example of such innovation is the elegant method suggested by Mishra *et al*. (2009) for disentangling the respective effects on

health of cumulative social disadvantage and exposure to low SEP at specific periods. This problem, identified earlier by Hallqvist *et al*. (2004), had previously appeared intractable.

Following previous practice, the main conclusions considered in this section are summarised in Box 19.4.1

BOX 19.4.1: Research Question 3 - statement of current knowledge, and what this study adds.

what is already known on this topic:

The sparse sampling schemes frequently adopted in inequality research (see Section 19.3) are convenient in terms of the relative ease with which they may be applied. Additional motivation for the continued use of such approaches is provided by their extensive record of success in predicting a range of health outcomes.

what this study adds:

Despite possible limitations identified in this study, widespread replacement of the traditional sparse sampling regime by more detailed (i.e. frequent) recording of social or environmental exposure is unlikely. Apart from practical considerations of cost and availability, this study suggests that the exploitation of 'dense' sequential data in analysis presents substantial challenges. These challenges mean that despite the richer information yielded by frequent sampling, its practical advantages over traditional two-, three- or four-point sampling are questionable.

Where specific circumstances justify the use of dense sampling, recent methodological developments such as semiparametric group-based methodology and functional data analysis offer scope for fully exploiting the rich data involved.

19.5 Concluding comments

In conclusion, some final comments relating to the generalisability of the findings reported and discussed above are offered. This study involved an in-depth examination of the social and residential experiences of 294 British individuals who were born between 1919 and 1934. As with any cohort, the experiences of these people were shaped by the social, cultural and political forces peculiar to their time. This is reflected in the data on which this study was based. For example, the high proportion of subjects reporting some form of Armed Forces affiliation indicates influences of the Second World War and the post-war era of compulsory National Service which ended in the 1960s. Similarly, as discussed earlier, the declining prevalence of exposure to air pollution over time may be attributable to specific legislative measures (the Clean Air Acts) which were enacted during subjects' lifetimes.

Because of this, the results of this study are limited in the extent to which they may be generalised to wider populations. Findings such as the bimodal distribution of cumulative disadvantage, or the heavily skewed distribution of exposure to residential dampness, might assume very different forms for a cohort born even ten years later.

Beyond the specific peculiarities of the cohort examined, the limitations of the study itself which were summarised in Section 19.2 provide further reasons for the exercise of caution in over-generalising its findings. A small study, arguably exhibiting certain inadequacies in both the data source used and the methods adopted, cannot safely be held up as an illustration of generally-applicable truths. Despite this, the main concern of the study (that is, the nature of the relationships linking social disadvantage, housing conditions and health) remains one which is of interest both to medical and social researchers, and to those who formulate social policy. There is a clear need for further research in this important but challenging area, and if nothing else the study has demonstrated some of the difficulties which will need to be surmounted if this research agenda is to be driven forward. Prominent among these is the formidable practical challenge of sourcing data with the required attributes; that is, detailed information on both social and residential exposures over a substantial portion of the lifecourse, paired with health outcomes captured in later life. However, even if such data are available, this study has suggested that their effective use in investigating associations of the kind examined here may face unexpected methodological difficulties.

Realistically, an adequate data source for investigating the model examined in this study is likely to become available only rarely. Because of this, future work in this field will probably continue the very extensive tradition of examining the individual associations involved (i.e. socioeconomic position with health, SEP with housing conditions, and housing conditions with health) separately, as distinct from attempting (as in this study) to assess the complete postulated causal chain. One specific strand of research which might be developed relatively easily is that of comparing the respective merits of 'sparse' and intensive sampling of socioeconomic position when assessing the influence of SEP on health. Existing research initiatives such as the SHEEP study (Hallqvist *et al*., 2004; Ljung & Hallqvist, 2006) have assembled data which permit the construction of detailed life-time trajectories of exposure to social disadvantage. This offers the opportunity to compare how effectively sparse and detailed representations of SEP predict specific health outcomes (in the case of the SHEEP study, myocardial infarction). Such investigations would potentially yield interesting new

insights into the validity of current assumptions about the near-ubiquity of the social patterning of health.

[END]

APPENDIX 1 *(six pages)*

Clustered representation of time-dependent socioeconomic position, sampled at yearly intervals in the age range from 15 to 60 years

LEGEND: 0 = manual, 1 = non manual, * = Armed Forces,
. = non-employed

Subject's status is shown at yearly intervals from 15 years (leftmost character in 'Sequence' string) to 60 years (rightmost character)

---------- CLUSTER=1 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
1	1111111111111111111111111111111111111111111111	0	6	6
2	1111111111111111111111100000111111111111111111	0	1	1
4	1111111111111111111100000000011111111111111111	0	1	1
5	1111111111100000000000111..1111111111111111	0	1	1
7	11111111100000111111111111111111111111111111100	0	1	1
10	111111..1111111111111111111111111111111111	1	0	1
13	111111****1111111111111111111111111111111111	1	0	1
15	1111..011111111111111111111111111111111....	1	0	1
17	111.1111111111111111111111111111111111111111	0	1	1
18	111..1111111111111111111111111111111111111	0	2	2
19	111..111..1111111111111111111111111111111111	0	1	1
20	111...1111111111111111111111111111111111111	0	2	2
21	111....1111111111111111111111111111111111111	0	3	3
22	111....1111111111111111111111111111111111.	1	0	1
23	111.....1111111111111111111111111111111111	1	2	3
24	111.....1111111....11111111111111111111111	0	1	1
25	111......1111111111111111111111111111111111	1	1	2
26	111.......1111111111111111111111111111111111	0	1	1
27	111..**111....1111111111111111111111111111	1	0	1
29	11101111111111111111111111111111111111111111	0	1	1
30	1110.11111111111111111111111111111111.....	1	0	1
31	1110.....1111111111111111111111111111111111	1	0	1
32	1110..000111111111111111111111111111111111.	1	0	1
33	11100011111111111111111111111111111111111111	2	1	3
36	11100001111111111111111111111111111111111111	0	1	1
48	111**11111111111111111111111111111111111111	1	0	1
49	111****11111111111111111111111111111111111	1	0	1
50	111*****11111111111111111111111111111111111	1	0	1
52	00011111111111111111111111111111111111111111	1	13	14
53	00011111111111111111111111111111111111.....	0	2	2
59	000111111111111.....11111111111111111111111	1	0	1
60	00011111111111000001100001111111111111111110	0	1	1
64	000111111.11111111111111111111111111111111	0	1	1
65	00011111100111111100000111111111111111111100	0	1	1
66	00011111100000111111111111111111111111111111	0	1	1
69	00011111.11111111111111111111111111111111	0	1	1
71	00011111000011111111111111111111111111111111	0	1	1
74	0001111.111111111111111111111111111111110	1	0	1
75	00011110000001111111111111111111111111111111	0	1	1

---------- CLUSTER=1 ----------
(continued)

Sequence ID	Sequence	No. Male	No. Female	Total No.
79	000111....11001111111111111111111111111111111	0	1	1
84	0001.11111111111111111111111111111111111111111	2	0	2
85	0001.11111111111111111111111001111111111111111	0	1	1
87	0001..11111111111111111111111111111111111111111	1	0	1
88	0001...11111111111111111111111111111111111111	0	1	1
91	000.111111111111111111111111111111111111111111	0	2	2
92	000...11111111111111111111111111111111111111	0	2	2
93	000...1111111111111111111111111111111111.....	1	0	1
94	000...11111111111111111111111000111111110011111	0	1	1
95	000.....111111111111111111111111111111111100	0	1	1
96	000......11111111111111111111111111111111111	1	0	1
97	000......11..11111111111111111111111111111111	1	0	1
98	000..........1111111111111111111111111111111	1	0	1
103	000....*****11111111111111111111111111111111	1	0	1
107	00001111111111111111111111111111111111111111111	1	0	1
108	00001111111111111111111111111111111000011111111	0	1	1
111	0000.11111111111111111111111111111111111111111	1	0	1
116	0000..01111111111111111111111111111111111111111	1	0	1
118	0000.011111111111111111111111111111111111111111	1	0	1
132	00000111111111111111111111111111111111111111111	1	0	1
134	0000011.1111111111111111111111111111111111111	1	0	1
136	00000...11111111111000000000111111111111110000	0	1	1
143	000000111111111111111111111111111111111111111111	0	1	1
148	0000000111111111111111111111111111111111111111111	0	1	1
153	00000000111111111111111111111111111111111111111	1	0	1
158	000000000111111111111111111111111111111111111111	0	1	1
198	000***11111111111111111111111111111111111111	1	0	1
208	000******...11111111111111111111111111111111	1	0	1
211	000*******1111111111111111111111111111111111	1	0	1
--------		----	------	-----
CLUSTER		36	63	99

---------- CLUSTER=2 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
12	11111100000000000000000000000000000000000000000	0	1	1
28	111.000000000000000000000000000000000000000000	0	1	1
35	111000.000000000000000000000000000000000000000	0	1	1
38	11100000000111000000000000000000000000000000000	0	1	1
40	11100000000...000000000000000000000000000000000	1	0	1
43	11100000000000000000000110000000000000000000000	1	0	1
47	11100000000000000000000000000000000000000000000	2	3	5
51	...000000000000000000000000000000000000000000.	1	0	1
63	00011111110000000000000000000000000000000000000	0	1	1
67	00011111100000000000000000000000000000000000000	0	2	2
77	0001111000000000000000000000000000000000000000	0	1	1
81	0001110000000000000000000000000000000000000000	0	1	1
83	0001100000000000000000000000000000000000000000	0	1	1
86	0001.1100000000000000000000000000000000000....	1	0	1
89	0001..0000000000000000000000000000000000000000	1	0	1

---------- CLUSTER=2 ----------
(continued)

Sequence ID	Sequence	No. Male	No. Female	Total No.
100	000.........0000000000000000000000000000000000	0	1	1
101	000......000000.0000000000000000000000000000000	1	0	1
102	000....00000000000000000000000000000000000000	0	1	1
105	000..00000000000000000000000000000000000000000	1	0	1
113	0000........1111000000000000000000000000000000	0	1	1
115	0000...000000000000000000000000000000000000000	2	0	2
117	0000..0000000000000000000000000000000000000000	1	0	1
125	0000.00000000000000000000000000000000........	1	0	1
126	0000.0000000000000000000000000000000000.....	1	0	1
127	0000.00000000000000000000000000000000000011111	1	0	1
128	0000.000000000000000000000000000000000000.....	1	0	1
129	0000.0000000000000000000000000000000000000....	1	0	1
130	0000.00000000000000000000000000000000000000..	1	0	1
131	0000.0000000000000000000000000000000000000000	5	0	5
133	0000011100000000000000000000000000000000000000	1	0	1
135	0000010000111100000000000000000000000000000000	0	1	1
138	00000..0000000000000000000000000010000000000000	1	0	1
140	00000.00000000000000000000000000000000000011111	1	0	1
141	00000.000000000000000000000000000000000000011.	1	0	1
142	00000.0000000000000000000000000000000000000000	1	0	1
147	000000.0000000000000000000000000000000000000000	0	1	1
149	0000000....0000000000000000000000000000000000000	1	0	1
154	00000000111000000000000000000000000000001111111	0	1	1
157	00000000.***0000000000000000000000000000000000	1	0	1
171	00000000000000000111110000000000000000000000000	0	1	1
172	00000000000000000100000000000000000000000000000	0	1	1
174	00000000000000000111111100000000000000000000000	1	0	1
175	00000000000000000111111100000000000000000000000	0	1	1
176	00000000000000000010000000000000000000000000000	0	1	1
179	00000000000000000000111100000000000000000000000	0	1	1
181	00000000000000000000000000.0000..0000000....	1	0	1
186	00000000000000000000000000000000000..........	1	0	1
187	00000000000000000000000000000000000111111111	0	1	1
188	00000000000000000000000000000000000........	1	0	1
189	00000000000000000000000000000000000000000.....	1	0	1
190	00000000000000000000000000000000000000000000...	1	0	1
191	00000000000000000000000000000000000000000000..	2	0	2
192	00000000000000000000000000000000000000000000001	0	1	1
193	00000000000000000000000000000000000000000000000.	1	0	1
194	00000000000000000000000000000000000000000000000	10	20	30
197	000**0000000000000000000000000000000000000000	2	1	3
199	000***100000.0000000000000000000000000000000000	1	0	1
203	000***0000000000000000000000000000000000000000	3	0	3
205	000****000000000000000000000000000000000000....	1	0	1
206	000*****00000011111111000000000000000000000000	1	0	1
207	000*****0000000000000000000000000000000000000	1	0	1
209	000******00000111111000000000000000000000000	1	0	1
210	000******0000000000000000000000000000000000000	2	0	2
212	000*******00000000000000000000000000000000000	1	0	1
CLUSTER		62	47	109

---------- CLUSTER=3 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
3	11111111111111111111111100000000000000000000000	0	1	1
11	11111100000011111111111111111110000000000000000	0	1	1
14	1111..11111111111100000111110000000000000000000	0	1	1
54	00011111111111111111111111110000000000000000000	0	1	1
55	00011111111111111111100000000000011110000000000	0	1	1
57	00011111111111111100000000000000000000000000000	0	1	1
58	00011111111111111000000000000000000000000000000	0	1	1
CLUSTER		0	7	7

---------- CLUSTER=4 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
34	11100011000000000000000001111111111111111111111	0	1	1
37	11100000011111111111111111111111.....	0	1	1
39	11100000000100111111111111111111111100011111111	0	1	1
41	11100000000000110111111111111111111111111111111	1	0	1
44	11100000000000000000000000111111111111111111111	1	0	1
61	000111111111.11111111111111111111110000000000	0	1	1
70	00011111000110000000000111111111111111111111111	0	1	1
76	00011110000000000000000001111110111111111111111	0	1	1
80	00011100000011111111111111111111111111000011111	0	1	1
82	000111**00000001111111111111111111111111.....	1	0	1
104	000..00000000000011111111111111111111110000000	1	0	1
109	00001111111111111111111100000000111...11111111	0	1	1
110	00001111111111111111111000001111111110000000000	0	1	1
119	0000.00000000001111111111111111111111111110000	1	0	1
120	0000.00000000000000011111111111111111111111111	1	0	1
121	0000.00000000000000001111111111111111111111111	1	0	1
122	0000.00000000000000000011111111111111111111111	1	0	1
139	00000.00000000111111111111111111111111111111111	1	0	1
144	00000011111111111111111111111111111000000000....	0	1	1
146	000000.00000000000000011111111111111111111111111	1	0	1
150	0000000.00001111111101111111111....111111111	1	0	1
151	0000000.00000000000000001111111111111111111111	1	0	1
156	00000000.00000011111111111111111111111111111111	1	0	1
159	00000000010001111111111111111111111111111111111	0	1	1
160	00000000001111111111111101111110000000001111111	0	1	1
161	00000000000111111111111111111111111111111111111	0	1	1
163	00000000000001111111111111111111111111111111111	0	1	1
164	00000000000001111111111111111111111111110000000	0	1	1
165	00000000000000111111111111111111111000000000000	0	1	1
166	0000000000000111..11111111111111111111111111	1	0	1
167	00000000000000011111111111111111111111111111111	2	0	2
169	0000000000000000...11111111111111111111111111	1	0	1
170	00000000000000000111111111111111111111111111111	0	2	2
173	00000000000000000011111111111111111111111111111	0	1	1
178	00000000000000000000011111111111111111111111111	0	1	1
180	00000000000000000000000111111111111111111111111	0	1	1
196	000**00000000000011111111111111111111111111111	1	0	1
201	000***0000000000000001111111111111111111111111	1	0	1
CLUSTER		19	21	40

---------- CLUSTER=5 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
8	11111111000000000000000000001111111111111100000	0	1	1
9	1111111000000000000000000011111111111111110000	0	1	1
16	1111..0000011100000000000111100011111111..1111	1	0	1
42	11100000000000000000111000001111111111111110000	0	1	1
45	11100000000000000000000000000000001111111111000	1	0	1
46	11100000000000000000000000000000000011111111111	0	1	1
56	00011111111111111100000000000000000111111111111	0	1	1
62	00011111111000000000000000000000111001111101111	0	1	1
72	00011111000000000000000000000000001111111100000	0	1	1
73	00011111000000000000000000000000000001111110000	0	1	1
112	0000..1111110000000000000000001111111111111111	1	0	1
123	0000.00000000000000000011111111111111111000000	1	0	1
124	0000.00000000000000000000000000011111111111111	1	0	1
145	000000......000.00000000000000000111111111111..	0	1	1
152	0000000.00000000000000000000000000011111111111	1	0	1
155	00000000..00000000000000001111111111111100000	1	0	1
162	00000000000111100000000001111111000000000.....	0	1	1
168	00000000000000011111110000000000000000001111111	0	1	1
177	00000000000000000000111111111111100000000000000	0	1	1
182	00000000000000000000000000000111111111111111111	1	0	1
183	00000000000000000000000000000111111111111011111	1	0	1
184	00000000000000000000000000000000111111111110001	0	1	1
185	00000000000000000000000000000000001111111111....	1	0	1
195	0000*****000000000000000000000000011111110000.	1	0	1
202	000***000000000000000000000000000111111111110	1	0	1
204	000****00000000000000000000111111100000.....	1	0	1
CLUSTER		13	13	26

---------- CLUSTER=6 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
78	0001111********************000000111111111111	0	1	1
90	0001000**********************1111111111111111	1	0	1
106	000.************************111111111111111111	1	0	1
114	0000....************************111111111111....	1	0	1
214	000************************1111111111111111..	1	0	1
CLUSTER		4	1	5

---------- CLUSTER=7 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
213	000********************11000000000000000000000	1	0	1
215	000*************************000000000000.0000	1	0	1
CLUSTER		2	0	2

---------- CLUSTER=8 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
137	00000.....00000000....111111111111111.........	1	0	1
200	000***0000000000000011111111111111111........	1	0	1
CLUSTER		2	0	2

---------- CLUSTER=9 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
99	000...................00000000000000..........	1	0	1
216	00***.....00000...000000000000000.............	0	1	1
CLUSTER		1	1	2

---------- CLUSTER=10 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
6	11111111111**	0	1	1
68	000111111**********1111111*******1111*********	0	1	1
CLUSTER		0	2	2
		139	155	294

[END of Appendix 1]

APPENDIX 2 *(four pages)*

Clustered representation of time-dependent socioeconomic position, sampled at five-year intervals in the age range from 15 to 60 years

LEGEND: 0 = manual, 1 = non manual, * = Armed Forces,
. = non-employed

Subject's status is shown at five-year intervals from 15 years (leftmost character in 'Sequence' string) to 60 years (rightmost character)

---------- Cluster=1 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
89	0111111111	7	18	25
115	1111111111	3	11	14
95	1.11111111	3	9	12
31	0.11111111	4	3	7
63	0011111111	2	3	5
105	1011111111	2	2	4
73	0101111111	0	3	3
94	1.1111111.	3	0	3
14	0*11111111	2	0	2
87	011111111.	0	2	2
91	1*11111111	2	0	2
113	1111101111	0	2	2
4	0*.1111111	1	0	1
16	0.*1111111	1	0	1
19	0..1111111	1	0	1
29	0.1111111.	1	0	1
30	0.11111110	0	1	1
74	011.111111	1	0	1
84	0111110111	0	1	1
85	0111111011	0	1	1
88	0111111110	1	0	1
90	1*.1111111	1	0	1
92	1.1.111111	0	1	1
111	1110011111	0	1	1
114	111111111.	1	0	1
--------		----	------	-----
cluster		36	58	94

---------- Cluster=2 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
38	0000000000	18	25	43

---------- Cluster=3 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
96	1000000000	4	6	10
6	0*00000000	8	0	8
37	000000000.	8	0	8
23	0.00000000	6	1	7
66	0100000000	1	5	6
36	00000000..	5	0	5
50	0000100000	1	2	3
39	0000000001	1	1	2
78	0111000000	0	2	2
1	.00000000.	1	0	1
5	0*0000000.	1	0	1
12	0*01000000	1	0	1
13	0*01100000	1	0	1
17	0..0000000	0	1	1
18	0..1000000	0	1	1
20	0.0.000000	1	0	1
22	0.0000000.	1	0	1
24	0.00000001	1	0	1
33	00*0000000	1	0	1
35	00.0000000	1	0	1
46	0000010000	0	1	1
65	010000000.	1	0	1
75	0110000000	0	1	1
106	1100000000	0	1	1
--------		----	------	-----
cluster		44	22	66

---------- Cluster=4 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
49	0000011111	5	2	7
59	0001111111	4	3	7
53	0000111111	2	3	5
69	0100011111	0	2	2
10	0*00011111	1	0	1
25	0.01111111	1	0	1
28	0.11110101	0	1	1
54	0001011.11	1	0	1
58	0001111110	1	0	1
61	001111100.	0	1	1
62	0011111011	0	1	1
70	0101011110	0	1	1
72	0101111101	0	1	1
77	0110111110	0	1	1
80	0111100.11	0	1	1
81	0111100100	0	1	1
82	0111101100	0	1	1
86	0111111100	0	1	1
100	1000011111	0	1	1
102	1000111111	1	0	1
103	1001111011	0	1	1

---------- Cluster=4 ----------
(continued)

Sequence ID	Sequence	No. Male	No. Female	Total No.
110	1101111110	0	1	1
cluster		16	23	39

---------- Cluster=5 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
2	0*****111.	1	0	1
3	0*****1111	1	0	1
15	0.****111.	1	0	1
32	00****1111	1	0	1
64	01****0111	0	1	1
cluster		4	1	5

---------- Cluster=6 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
42	0000000111	2	1	3
7	0*0000011.	1	0	1
8	0*00000110	1	0	1
9	0*0000110.	1	0	1
34	00..00011.	0	1	1
40	0000000011	0	1	1
41	000000011.	1	0	1
43	0000001101	1	0	1
44	0000001110	1	0	1
45	0000001111	1	0	1
47	00000110..	0	1	1
55	0001100011	0	1	1
60	0010000011	0	1	1
67	0100000010	0	1	1
68	0100000110	0	1	1
76	0110000011	0	1	1
97	1000000110	1	0	1
98	1000000111	0	1	1
99	1000001111	1	0	1
cluster		11	10	21

---------- Cluster=7 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
26	0.10000111	1	0	1
27	0.11001110	0	1	1
48	0000011100	1	0	1
51	0000111000	0	1	1
52	0000111100	1	0	1
56	0001111000	0	1	1
57	0001111100	0	1	1
79	0111000111	0	1	1
101	1000101110	0	1	1
107	1100001110	0	1	1
108	1100011110	0	1	1
cluster		3	8	11

---------- Cluster=8 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
11	0*001111..	1	0	1
21	0.00.111..	1	0	1
71	01011111..	1	0	1
104	1011111...	0	1	1
cluster		3	1	4

---------- Cluster=9 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
83	0111110000	0	1	1
93	1.11010000	0	1	1
109	1101110000	0	1	1
112	1111100000	0	1	1
cluster		0	4	4
		135	152	287

[END of Appendix 2]

APPENDIX 3 *(two pages)*

Clustered representation of time-dependent exposure to residential dampness, sampled at five-year intervals in the age range from 15 to 60 years

LEGEND: 0 = not exposed, 1 = exposed

Subject's status is shown at five-year intervals from 15 years (leftmost character in 'Sequence' string) to 60 years (rightmost character)

---------- Cluster=1 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
1	0000000000	53	76	129

---------- Cluster=2 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
18	0010000000	8	4	12
13	0001000000	3	3	6
30	1000000000	2	2	4
3	0000000100	0	3	3
9	0000100000	0	3	3
14	0001100000	1	2	3
6	0000010000	1	1	2
20	0011000000	1	1	2
21	0011100000	2	0	2
25	0100000000	1	1	2
2	0000000011	0	1	1
4	0000001000	1	0	1
10	0000110000	0	1	1
--------		----	------	-----
cluster		20	22	42

---------- Cluster=3 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
12	0000111111	1	3	4
17	0001111111	2	1	3
11	0000111100	1	1	2
15	0001111000	0	2	2
16	0001111100	0	2	2
23	0011111100	2	0	2
22	0011111000	1	0	1
--------		----	------	-----
cluster		7	9	16

---------- Cluster=4 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
31	1100000000	2	3	5
26	0110000000	1	1	2
27	0111000000	1	0	1
28	0111100000	0	1	1
32	1100000100	0	1	1
35	1110000000	0	1	1
36	1110000001	1	0	1
37	1110000010	1	0	1
39	1111000000	1	0	1
cluster		7	7	14

---------- Cluster=5 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
24	0011111111	0	2	2
29	0111111111	0	2	2
41	1111111111	0	2	2
38	1110111111	1	0	1
40	1111111100	1	0	1
cluster		2	6	8

---------- Cluster=6 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
5	0000001111	1	2	3
33	1100001110	2	0	2
7	0000011100	1	0	1
8	0000011111	1	0	1
19	0010001111	1	0	1
34	1100001111	1	0	1
cluster		7	2	9
		96	122	218

[END of Appendix 3]

APPENDIX 4 *(three pages)*

Clustered representation of time-dependent exposure to air pollution, sampled at five-year intervals in the age range from 15 to 60 years

LEGEND: 0 = not exposed, 1 = exposed

Subject's status is shown at five-year intervals from 15 years (leftmost character in 'Sequence' string) to 60 years (rightmost character)

---------- Cluster=1 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
1	0000000000	59	66	125

---------- Cluster=2 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
10	0010000000	1	3	4
20	1000000000	1	2	3
7	0001000000	1	1	2
14	0100000000	0	2	2
2	0000000001	0	1	1
4	0000010000	0	1	1
6	0000110000	0	1	1
8	0001100000	1	0	1
11	0010010000	1	0	1
15	0100110000	0	1	1
21	1000010000	1	0	1
22	1001000000	1	0	1
--------		----	------	-----
cluster		7	12	19

---------- Cluster=3 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
26	1110000000	9	7	16
24	1100000000	5	6	11
16	0110000000	0	2	2
17	0111000000	1	0	1
23	1010000000	0	1	1
25	1100100000	0	1	1
--------		----	------	-----
cluster		15	17	32

---------- Cluster=4 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
28	1111000000	3	1	4
30	1111100000	1	2	3
31	1111100001	0	2	2
32	1111110000	2	0	2
27	1110011000	0	1	1
cluster		6	6	12

---------- Cluster=5 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
37	1111111111	3	7	10
36	1111111110	1	1	2
19	0111111111	0	1	1
29	1111011111	0	1	1
33	1111110011	1	0	1
cluster		5	10	15

---------- Cluster=6 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
3	0000001111	2	0	2
5	0000011111	1	0	1
cluster		3	0	3

---------- Cluster=7 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
9	0001111111	0	1	1
12	0011100111	0	1	1
13	0011111111	0	1	1
cluster		0	3	3

---------- Cluster=8 ----------

Sequence ID	Sequence	No. Male	No. Female	Total No.
34	1111111000	0	2	2
18	0111111000	1	0	1
35	1111111100	0	1	1
cluster		1	3	4
		96	117	213

[END of Appendix 4]

APPENDIX 5 *(five pages)*

Detailed sequences of socioeconomic position, sampled at individual years in the age range from 15 to 60

LEGEND: 0 = manual, 1 = non manual, * = Armed Forces,
. = non-employed

Subject's status is shown at yearly intervals from 15 years (leftmost character in 'Sequence' string) to 60 years (rightmost character)

Sequence ID	Sequence	No. of subjects
1	000000000000000000000000000000000000000000000000	30
2	000111111111111111111111111111111111111111111111	14
3	111111111111111111111111111111111111111111111111	6
4	0000.000000000000000000000000000000000000000000	5
5	111000000000000000000000000000000000000000000000	5
6	000***000000000000000000000000000000000000000000	3
7	000**000000000000000000000000000000000000000000	3
8	111.....111111111111111111111111111111111111111	3
9	111....111111111111111111111111111111111111111	3
10	111000111111111111111111111111111111111111111111	3
11	000******000000000000000000000000000000000000	2
12	000...111111111111111111111111111111111111111	2
13	000.111111111111111111111111111111111111111111	2
14	0000...000000000000000000000000000000000000000	2
15	000000000000000000000000000000000000000000..	2
16	000000000000000001111111111111111111111111111	2
17	000000000000000111111111111111111111111111111	2
18	0001.111111111111111111111111111111111111111111	2
19	000111111000000000000000000000000000000000000	2
20	000111111111111111111111111111111111111111.....	2
21	111......111111111111111111111111111111111111	2
22	111...111111111111111111111111111111111111111	2
23	111..111111111111111111111111111111111111111	2
24	...000000000000000000000000000000000000000.	1
25	00***.....00000...000000000000000.....	1
26	000*************************000000000000.0000	1
27	000***********************1111111111111111..	1
28	000********************11000000000000000000	1
29	000*******000000000000000000000000000000000	1
30	000*******111111111111111111111111111111111	1
31	000******...111111111111111111111111111111	1
32	000******000001111110000000000000000000000000	1
33	000*****000000000000000000000000000000000000	1
34	000*****000000111111100000000000000000000000	1
35	000****000000000000000000000000000000000....	1
36	000****000000000000000000001111111000000.....	1
37	000***000000000000000000000000001111111111110	1
38	000***000000000000000111111111111111111111111	1
39	000***000000000000001111111111111111........	1
40	000***100000.000000000000000000000000000000	1
41	000***111111111111111111111111111111111111111	1
42	000**000000000001111111111111111111111111111	1
43	000.**********************111111111111111111	1

Sequence ID	Sequence	No. of subjects
44	000....*****1111111111111111111111111111111111	1
45	000..................00000000000000..........	1
46	000..........11111111111111111111111111111111	1
47	000.........00000000000000000000000000000000	1
48	000......000000.00000000000000000000000000000	1
49	000......11..11111111111111111111111111111111	1
50	000......11111111111111111111111111111111111	1
51	000.....111111111111111111111111111111111100	1
52	000....00000000000000000000000000000000000000	1
53	000...11111111111111111111111000111111100111111	1
54	000...11111111111111111111111111111111111.....	1
55	000..00000000000000000000000000000000000000000	1
56	000..00000000000111111111111111111111110000000	1
57	0000*****00000000000000000000000011111110000.	1
58	0000....**********************111111111111....	1
59	0000........11110000000000000000000000000000	1
60	0000..00000000000000000000000000000000000000000	1
61	0000..01111111111111111111111111111111111111111	1
62	0000..11111100000000000000000011111111111111111	1
63	0000.00000000000000000000000000000000........	1
64	0000.000000000000000000000000000000000.......	1
65	0000.0000000000000000000000000000000000.....	1
66	0000.00000000000000000000000000000000000....	1
67	0000.000000000000000000000000000000000000..	1
68	0000.0000000000000000000000000000000000011111	1
69	0000.0000000000000000000000000011111111111111	1
70	0000.0000000000000000001111111111111111000000	1
71	0000.0000000000000000011111111111111111111111	1
72	0000.0000000000000001111111111111111111111111	1
73	0000.0000000000000011111111111111111111111111	1
74	0000.0000000001111111111111111111111110000	1
75	0000.01111111111111111111111111111111111111111	1
76	0000.11111111111111111111111111111111111111111	1
77	00000.....00000000....111111111111111.........	1
78	00000...11111111111000000000111111111111110000	1
79	00000..00000000000000000000000001000000000000	1
80	00000.000000000000000000000000000000000000000	1
81	00000.0000000000000000000000000000000000011.	1
82	00000.0000000000000000000000000000000000011111	1
83	00000.0000000011111111111111111111111111111111	1
84	000000......000.00000000000000000111111111111..	1
85	000000.000000000000000000000000000000000000000	1
86	000000.0000000000000001111111111111111111111111	1
87	0000000....000000000000000000000000000000000000	1
88	0000000.0000000000000000000000000011111111111	1
89	0000000.0000000000000001111111111111111111111	1
90	0000000.000011111111101111111111111....111111111	1
91	00000000.***000000000000000000000000000000000000	1
92	00000000..0000000000000001111111111111111100000	1
93	00000000.000000111111111111111111111111111111111	1
94	0000000000000000...11111111111111111111111111	1
95	00000000000000000000000000000.0000..0000000....	1
96	000000000000000000000000000000000000..........	1
97	000000000000000000000000000000000000000........	1
98	000000000000000000000000000000000000000000.....	1

Sequence ID	Sequence	No. of subjects
99	000000000000000000000000000000000000000000...	1
100	000000000000000000000000000000000000000000.	1
101	000000000000000000000000000000000000000000001	1
102	000000000000000000000000000000000000000111111111	1
103	000000000000000000000000000000001111111111....	1
104	000000000000000000000000000000111111111110001	1
105	000000000000000000000000000001111111111011111	1
106	000000000000000000000000000001111111111111111	1
107	000000000000000000000001111111111111111111111	1
108	000000000000000000000111100000000000000000000	1
109	000000000000000000000111111111111111111111111	1
110	000000000000000000001111111111111000000000000	1
111	000000000000000000100000000000000000000000000	1
112	000000000000000001111110000000000000000000000	1
113	000000000000000001111111000000000000000000000	1
114	000000000000000001111111111111111111111111111	1
115	000000000000000010000000000000000000000000000	1
116	000000000000000011111000000000000000000000000	1
117	000000000000001111111000000000000000001111111	1
118	0000000000000111..111111111111111111111111111	1
119	000000000000011111111111111111111000000000000	1
120	000000000000111111111111111111111111110000000	1
121	000000000000111111111111111111111111111111111	1
122	000000000011110000000000111111100000000......	1
123	000000000011111111111111111111111111111111111	1
124	000000000111111111111110111111000000000111111	1
125	000000001000111111111111111111111111111111111	1
126	000000001111111111111111111111111111111111111	1
127	000000011100000000000000000000000000001111111	1
128	000000011111111111111111111111111111111111111	1
129	000000111111111111111111111111111111111111111	1
130	000001111111111111111111111111111000000000....	1
131	000001111111111111111111111111111111111111111	1
132	000010000111100000000000000000000000000000000	1
133	0000011.111111111111111111111111111111111111	1
134	000001110000000000000000000000000000000000000	1
135	000001111111111111111111111111111111111111111	1
136	000011111111111111111110000011111111000000000	1
137	000011111111111111111111000000000111...11111111	1
138	000011111111111111111111111111111110000111111111	1
139	000011111111111111111111111111111111111111111111	1
140	0001...111111111111111111111111111111111111111	1
141	0001..000000000000000000000000000000000000000000	1
142	0001..111111111111111111111111111111111111111111	1
143	0001.110000000000000000000000000000000000....	1
144	0001.111111111111111111111110011111111111111111	1
145	0001000**********************1111111111111111	1
146	000110000000000000000000000000000000000000000000	1
147	000111**000000011111111111111111111111111......	1
148	000111....110011111111111111111111111111111111	1
149	000111000000000000000000000000000000000000000000	1
150	000111000000111111111111111111111111000011111	1
151	0001111********************000000111111111111	1
152	0001111.111111111111111111111111111111111110	1
153	000111100000000000000000000000000000000000000000	1
154	000111100000000000000000111111011111111111111	1

Sequence ID	Sequence	No. of subjects
155	000111100000011111111111111111111111111111111111	1
156	00011111.11111111111111111111111111111111111111	1
157	000111110000000000000000000000000000001111110000	1
158	000111110000000000000000000000000001111111100000	1
159	000111110000111111111111111111111111111111111111	1
160	000111110001100000000001111111111111111111111111	1
161	00011111**********1111111*******1111*********	1
162	000111111.11111111111111111111111111111111111111	1
163	000111111000001111111111111111111111111111111111	1
164	000111111001111111000001111111111111111111111100	1
165	000111111100000000000000000000000000000000000000	1
166	000111111110000000000000000000111001111101111	1
167	000111111111.11111111111111111111110000000000	1
168	000111111111110000011000011111111111111111111110	1
169	00011111111111.....11111111111111111111111111	1
170	000111111111111110000000000000000000000000000000	1
171	000111111111111111000000000000000000000000000000	1
172	000111111111111111000000000000000111111111111	1
173	000111111111111111110000000000011110000000000	1
174	000111111111111111111111110000000000000000000000	1
175	111*****11111111111111111111111111111111111111	1
176	111****11111111111111111111111111111111111111	1
177	111**11111111111111111111111111111111111111	1
178	111..**111....11111111111111111111111111111111	1
179	111.......11111111111111111111111111111111111	1
180	111.....1111111....11111111111111111111111111	1
181	111....11111111111111111111111111111111111.	1
182	111..111..11111111111111111111111111111111111	1
183	111.000000000000000000000000000000000000000000	1
184	111.11111111111111111111111111111111111111111	1
185	1110.....11111111111111111111111111111111111	1
186	1110..000111111111111111111111111111111111.	1
187	1110.11111111111111111111111111111111.....	1
188	111000.000000000000000000000000000000000000	1
189	11100000000..000000000000000000000000000000	1
190	11100000000000000000000000000000011111111111	1
191	111000000000000000000000000000001111111111000	1
192	11100000000000000000000001111111111111111111	1
193	11100000000000000000011000000000000000000000	1
194	11100000000000000011100000111111111111110000	1
195	11100000000001101111111111111111111111111111	1
196	11100000000100111111111111111111111000111111	1
197	11100000000111000000000000000000000000000000	1
198	11100000011111111111111111111111.....	1
199	11100001111111111111111111111111111111111111	1
200	11100011000000000000000011111111111111111111	1
201	11101111111111111111111111111111111111111111	1
202	1111..000001110000000000111100011111111..1111	1
203	1111..011111111111111111111111111111111....	1
204	1111..11111111111000001111100000000000000000	1
205	111111****11111111111111111111111111111111111	1
206	111111..11111111111111111111111111111111111	1
207	11111100000000000000000000000000000000000000	1
208	11111100000011111111111111111000000000000000	1
209	11111110000000000000000011111111111111110000	1

Sequence ID	Sequence	No. of subjects
210	11111111000000000000000000011111111111100000	1
211	11111111100000111111111111111111111111111100	1
212	1111111111**********************************	1
213	11111111111100000000000111..1111111111111111	1
214	11111111111111111111100000000011111111111111	1
215	11111111111111111111111100000000000000000000	1
216	11111111111111111111111100000111111111111111	1
		294

[END of Appendix 5]

APPENDIX 6 *(three pages)*

Detailed sequences of exposure to residential dampness, sampled at individual years in the age range from 15 to 60

LEGEND: 0 = not exposed, 1 = exposed

Subject's status is shown at yearly intervals from 15 years (leftmost character in 'Sequence' string) to 60 years (rightmost character)

Sequence ID	Sequence	No. of subjects
1	000000000000000000000000000000000000000000000000	116
2	000000000000000111111111111111111111111111111111	2
3	000000000000001110000000000000000000000000000000	2
4	000000000001110000000000000000000000000000000000	2
5	000000000011000000000000000000000000000000000000	2
6	000000001111100000000000000000000000000000000000	2
7	111111110000000000000000000000000000000000000000	2
8	111111111100000000000000000000000000000000000000	2
9	000000000000000000000000000000000000001111111111	1
10	000000000000000000000000000000000000011110000000	1
11	000000000000000000000000000000001111111110000000	1
12	000000000000000000000000000000001111111111111111	1
13	000000000000000000000000000001111111111111111111	1
14	000000000000000000000000000011100000000000000000	1
15	000000000000000000000000000111000000000000000000	1
16	000000000000000000000000001111111111111111111111	1
17	000000000000000000000001100000000000000000000000	1
18	000000000000000000000011111111111111111111111111	1
19	000000000000000000000100000000000000000000000000	1
20	000000000000000000000111111100000000000000000000	1
21	000000000000000000001111111111111111111111111111	1
22	000000000000000000011100000000000000000000000000	1
23	000000000000000000011111111111111111111111111111	1
24	000000000000000001111111111000000000000000000000	1
25	000000000000000001111111111111111111110000000000	1
26	000000000000000001111111111111111111111111111111	1
27	000000000000000011111100000000000000000000000000	1
28	000000000000000011111111111111111111110000000000	1
29	000000000000000011111111111111111111111111111111	1
30	000000000000000110000000000000000000000000000000	1
31	000000000000000110000000000000000000000000000000	1
32	000000000000000111111110000000000000000000000000	1
33	000000000000001111100000000000000000000000000000	1
34	000000000000001111111111000000000000000000000000	1
35	000000000000001111111111111111110000000000000000	1
36	000000000000011000000000000000000000000000000000	1
37	000000000000011000000001111111111111100000000000	1
38	000000000000011111000000000000000000000000000000	1
39	000000000000011111110000000000000000000000000000	1
40	000000000000011111111111111111110000000000000000	1
41	000000000000011111111111111111111111111111111111	1
42	000000000000111000000000000000000000000000000000	1
43	000000000000111111111100000000000000000000000000	1
44	000000000001000000000000000000000000000000000000	1

Sequence ID	Sequence	No. of subjects
45	000000000001100000000000000000000000000000000000	1
46	000000000001111111111111111111111111110000000000	1
47	000000000011111111110000000000000000000000000000	1
48	000000000011111111111111100000000000000000000000	1
49	000000000011111111111111111111111111111111111111	1
50	000000000110000000000000000000000000000000000000	1
51	000000000111000000000000000000000000000000000000	1
52	000000000111100000000000000000000000000000000000	1
53	000000000111110000000000000000000000000000000000	1
54	000000000111111111111110000000000000000000000000	1
55	000000000111111111111111111111111111111110000000	1
56	000000001001111000011000000000000000000000000000	1
57	000000001100000000000000000000000000000000000000	1
58	000000001111000000000000001111111111111111111111	1
59	000000001111111000000000000000000000000000000000	1
60	000000001111111111111111111111111111100000000000	1
61	000000011111000000000000000000000000000000000000	1
62	000000011111100000000000000000000000000000000000	1
63	000000011111111111110000000000000000000000000000	1
64	000000011111111111111111111111111111110000000000	1
65	000000011111111111111111111111111111111111111111	1
66	000000100000000000000000000000000000000000000000	1
67	000000100000111111111111111111111111000000000000	1
68	000000111000000000000000000000000000000000000000	1
69	000000111110000000000000000000000000000000000000	1
70	000001111000000000000000000000000000000000000000	1
71	000001111110000000000000001111000000000000000000	1
72	000001111111100000000000000000000000000000000000	1
73	000001111111111111000111111111111111100000000000	1
74	000001111111111111111111111111111111111111111111	1
75	000010000000000000000000000000000000000000000000	1
76	000110000000000000000000000001111111000000000000	1
77	001111111111111111111000000000000000000000000000	1
78	011111000000000000001110000000000000000000000000	1
79	011111000011111111111111111111111111111111111111	1
80	011111111111111100000000000000000000000000000000	1
81	100000000000000000000000000000000000000000000000	1
82	111000000000000000000000000000000000000000000000	1
83	111100000000000000000000000000000000000000000000	1
84	111110000000000000000000000000000000110000000000	1
85	111111000000000000000000000000000011000000000000	1
86	111111100000000000000000000000000000000000000000	1
87	111111110000000000000000001111111111111110000	1
88	111111111000000000000000000010000111111111100	1
89	111111111000000000000000011111111111111111111	1
90	111111111111000000000000000000000000000000000000	1
91	111111111111100000000000000000000000011111000	1
92	111111111111100111111111111111111111111111111111	1
93	111111111111111000000000000000000000000000000001	1
94	111111111111111000011111111111111111111111111111	1
95	111111111111111110000000000000000000000000000000	1

Sequence ID	Sequence	No. of subjects
96	11111111111111111111111111000000000000000000000	1
97	11111111111111111111111111111111111111110000000	1
98	11111111111111111111111111111111111111111111111	1
		========
		220

[END of Appendix 6]

APPENDIX 7 *(two pages)*

Detailed sequences of exposure to air pollution, sampled at individual years in the age range from 15 to 60

LEGEND: 0 = not exposed, 1 = exposed

Subject's status is shown at yearly intervals from 15 years (leftmost character in 'Sequence' string) to 60 years (rightmost character)

Sequence ID	Sequence	No. of subjects
1	000000000000000000000000000000000000000000000000	118
2	111111111111111111111111111111111111111111111111	8
3	111111111000000000000000000000000000000000000000	6
4	111111111111111000000000000000000000000000000000	5
5	111111100000000000000000000000000000000000000000	4
6	111111111111100000000000000000000000000000000000	4
7	111111111110000000000000000000000000000000000000	3
8	111111111111000000000000000000000000000000000000	3
9	111111111111111111100000000000000000000000000000	2
10	000000000000000000000000000000000000000011111	1
11	000000000000000000000000000011111111111111111	1
12	000000000000000000000000001111111111111111111	1
13	000000000000000000000011110000000000000000000	1
14	000000000000000000000100000000000000000000000	1
15	000000000000000000001111111111111111111111111	1
16	000000000000000001111111111000000000000000000	1
17	000000000000000111100000000000000000000000000	1
18	000000000000001111111111111111111111111111111	1
19	000000000000011111110000000000000000000000000	1
20	000000000000111000000000000000000000000000000	1
21	000000000000011111110000000000000000000000000	1
22	000000000000011111111111111111000000000000000	1
23	000000000100000000000000000000000000000000000	1
24	000000000111000000000000000000000000000000000	1
25	000000001000000000000000000000000000000000000	1
26	000000001111000000000000000000000000000000000	1
27	000000001111111000000000000000000000000000000	1
28	000000011000000000000000000000000000000000000	1
29	000000011100000000000011000000000000000000000	1
30	000000011111111111111000000000011111111111111	1
31	000000100011100000000000000000000000000000000	1
32	000000111000000000000000000000000000000000000	1
33	000001111111111111111111111111111111111111111	1
34	000011111111111000000000000000000000000000000	1
35	000011111111111110000000000000000000000000000	1
36	000011111111111111111111111111000000000000000	1
37	000011111111111111111111111111111111111111111	1
38	000110000000000000000000000000000000000000000	1
39	000111000000000000000000000000000000000000000	1
40	000111000000000011111000111111000000000000000	1
41	000111111100000000000000000000000000000000000	1
42	000111111111000000000000000000000000000000000	1
43	100000000000000000000000000000000000000000000	1
44	100000000000000000000011110000000000000000000	1

Sequence ID	Sequence	No. of subjects
45	11000000000000000000000000000000000000000000000000	1
46	11100000000000000000000000000000000000000000000000	1
47	11100000011111000000000000000000000000000000000000	1
48	11111000000001111110000000000000000000000000000000	1
49	11111100000000000000000000000000000000000000000000	1
50	11111100000000000000001111111111111110000000000	1
51	11111100111110000000000000000000000000000000000000	1
52	11111111000000000001111000000000000000000000000000	1
53	11111111011111111111111111111111111111111111111111	1
54	11111111111011111000000000000000000000000000000000	1
55	11111111111111100000000011111111000000000000000000	1
56	11111111111111111110000011111111111111111111111111	1
57	11111111111111111111000000000000000000000000000000	1
58	11111111111111111111000000000000000000000000000111	1
59	11111111111111111111100000000000000000000000000000	1
60	11111111111111111111110000000000000000000000000000	1
61	11111111111111111111110000000000000000000001111	1
62	11111111111111111111111100000000000000000000000000	1
63	11111111111111111111111110000000000000000000000000	1
64	11111111111111111111111111000000000000000000000000	1
65	11111111111111111111111111000000000000001111111	1
66	11111111111111111111111111111000000000000000000000	1
67	11111111111111111111111111111000011111111111111	1
68	11111111111111111111111111111111100000000000000	1
69	11111111111111111111111111111111111000000000000	1
70	11111111111111111111111111111111111111100000	1
71	11111111111111111111111111111111111111110000	1
		========
		215

[END of Appendix 7]

APPENDIX 8

Abstract of presentation given at the 52nd Annual Scientific Meeting of the Society for Social Medicine (17-19 September 2008)

REPRESENTING TRAJECTORIES OF SOCIAL LOCATION OVER THE LIFECOURSE: A CHALLENGE AND A PROPOSED SOLUTION

J Walker, R Mitchell, S Platt, D Blane.

Background: The near-ubiquity of social inequalities in health has long been recognised and extensively investigated. However, many studies in this field have been based on data that represent the individual's socioeconomic status at a single point in time (eg, adulthood) or at a small number of widely spaced discrete time points (eg, childhood, early adulthood and late adulthood). Such ''sparse'' sampling may mask changes or fluctuations in subjects' social position that occur between sampling points, potentially leading to the associations between a person's true social location over time and health being incorrectly estimated.

Objectives: To illustrate the diversity and complexity of individuals' trajectories of social location over the age range 15–60 years. To propose a method for reducing the diverse range of observed trajectories into a smaller number of higher-level patterns, thus facilitating the investigation of associations between social location and health status in later life.

Design: Retrospective longitudinal study.

Setting: United Kingdom.

Population: Men and women aged 63–78 years.

Results: Among the 294 individuals studied, a total of 216 unique trajectories were identified when the subject's social location at each individual year in the age range 15–60 years was defined as either manual, non-manual, unemployed or engaged in Armed Forces service. When sampling was restricted to every fifth year in the age range 15–60 years (10 datum points), a total of 122 unique trajectories of time-related social position was observed. Reducing the state space (by combining the manual and Armed Forces states) resulted in a small reduction in the number of trajectories identified. A data reduction process based on optimal matching (to derive statistical distances between trajectories) followed by cluster analysis (to identify natural groupings in the trajectories) was developed. Application of this process reduced the 122 unique 5-year trajectories to 13 clusters. However, interpretation of the resulting cluster scheme was problematical.

Conclusions: Life-time trajectories of social position, when examined at the single or 5-year level, exhibit considerable diversity and variety. Such detailed variation may not be accurately captured in studies that consider a person's social status only at a small number of discrete time points. Current interest in life-course explanations of health inequalities may demand more detailed representations of subjects' life-time social position than have generally been used hitherto. Further research is required into techniques for reducing large sets of unique social trajectories into more compact higher-level schemes, to permit effective analysis.

(reproduced from Walker et al*. [2008], p. A2)*

NOTE ON LITERATURE SEARCHING

This note provides a very brief outline of the general approach which was adopted to locate literature relevant to the topic covered by the study.

Following formulation of the study's broad objectives, two general themes were identified as being of major importance: social inequality in health (that is, the nature of relationships between SEP and health), and the health impacts of the domestic environment. The author of this thesis had some familiarity with the latter through his previous research activities (Walker *et al*., 2006; Walker *et al*., 2009). Consequently, the task of isolating a core of the most influential literature relating to this sub-topic was relatively undemanding, in that many relevant references were already to hand. However, material with which the author was familiar was augmented by initially conducting searches on MEDLINE using the arguments listed below. In each case, the results were limited to review articles in the English language, published between 1979 and 2009. The main arguments used were:-

hous* AND asthma (296 citations located)
hous* and cardiovascular (59)
hous* AND stroke (20)
hous* AND myocard* (22)
hous* AND resp* (890)

Because of the study's concentration on the specific residential hazard of dampness (and its manifestation via the presence of moulds / fungi), these searches were repeated with the first term replaced to reflect these environmental features e.g.

damp* AND asthma
(mold* OR mould*) AND resp*

On completion of this process, the results of all searches were logically ORed (to eliminate duplicates), yielding a single set of review papers covering the area of interest. These were subjected to a manual scrutiny of titles, which immediately excluded those results appearing to be of no direct relevance to the study. For example, a paper such as that by Yoder *et al*. (2008) entitled *The giant Madagascar hissing-cockroach* (Gromphadorhina portentosa) *as a source of antagonistic moulds: concerns arising from its use in a public setting* seemed

unlikely to be of interest[56]. Further results were eliminated following scrutiny of abstracts. Of those eliminated in this latter way, the most common reason for exclusion was that they had a highly technical biomedical focus. An example is a review by Reed & Kita (2004) of the role of proteases [enzymes which break down proteins] in the activation of inflammation in allergic respiratory diseases.

A broadly similar series of searches was conducted using the Social Sciences Citation Index (SSCI). Results from the two series were combined, and the set of review articles thus obtained served as the starting point for the identification of relevant literature in the field of housing conditions and health. As is almost always the case when reviewing academic literature, the process was organic: paper *A* might cite sources *B, C* and *D*, each of which in turn suggested further reading, thus driving the progressive accumulation of a large body of relevant material.

The approach adopted to identify literature relevant to the other main sub-theme (social inequality in health) was somewhat different. Here, the starting point was not an initial literature search, but a small number of seminal texts which were recommended at an early stage by the author's supervisors and other colleagues. Specifically, a reading of Whitehead (1992), Kuh & Ben-Shlomo (1997), Shaw *et al*. (1999), Davey Smith *et al*. (2001) and Bartley (2004) identified a very large number of relevant references, including many which were central to the study's objectives. As before, investigating these sources led organically to the identification of further literature. Recent developments in the area of health inequality were monitored by regular reference to the contents tables and advance access facilities of appropriate journals, notably the *International Journal of Epidemiology*, the *Journal of Epidemiology and Community Health* and the *Journal of Public Health.* Similar attention was paid to leading medical journals such as the *BMJ* and the *Lancet* which, although primarily vehicles for the propagation of clinical knowledge, carry many papers which recognise socioeconomic aspects of health and health care.

Searching of literature databases was employed selectively, using fairly complex search arguments tailored to specific areas of interest. For example, location of material potentially

[56] In fact, titles can be deceptive. Had the study been based on data covering a sample of American individuals, this paper might well have been retained: these creatures are seemingly popular as children's pets, and are often reared in US schools for educational purposes. The authors concluded that "Cockroach rearing conditions thus serve as an ideal environment for mould growth and proliferation, and the subsequent use (handling) of these insects in a public forum increases the risk of inducing mould-related allergies in humans." (Yoder *et al*. 2008, p. 95)

relevant to the psychosocial model of inequality in the context of cardiovascular disease was accomplished by searching MEDLINE using the argument

> (psychosocial OR psycho-social) AND
> (socioeconomic OR socio-economic)
> AND cardio*).

This tailored use of database searching was in contrast to the approach adopted (as described earlier) when examining the literature on housing conditions and health. Searches applied in connection with the latter tended to be more general in scope.

A third strand of literature, not covered so far, was purely methodological. This included material relating to, *inter alia*, sequence (string) comparison and cluster analysis. In general, such literature was not located via extensive searching of electronic databases. Rather, reliance was placed on the author's pre-existing knowledge of these topics, augmented by reading a small number of key texts (such as Sankoff & Kruskal's [1999] seminal contribution to the field of sequence comparison).

REFERENCES

Abbott, A. & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History* 16, 471-494.

Abbott, A. & Hrycak, A. (1990). Measuring resemblance in sequence data: an optimal matching analysis of musicians' careers. *American Journal of Sociology* 96, 144-185.

Abbott, A. & DeViney, S. (1992). The welfare state as transnational event: evidence from sequences of policy adoption. *Social Science History* 16, 245-274.

Adams, J., White, M., Pearce, M. S., & Parker, L. (2004). Life course measures of socioeconomic position and self reported health at age 50: prospective cohort study. *Journal of Epidemiology & Community Health* 58, 1028-1029.

Albert, A. & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71, 1-10.

Altman, D. G. (1991). *Practical Statistics for Medical Research*. London, Chapman & Hall.

Arlian, L. G. & Platts-Mills, T. A. E. (2001). The biology of dust mites and the remediation of mite allergens in allergic disease. *Journal of Allergy and Clinical Immunology* 107, S406-S413.

Armitage, P. & Berry, G. (1987). *Statistical Methods in Medical Research*, 2nd ed. Oxford, Blackwell Scientific Publications.

Avendano, M., Glymour, M. M., Banks, J., & Mackenbach, J. P. (2009). Health disadvantage in US adults aged 50 to 74 years: a comparison of the health of rich and poor Americans with that of Europeans. *American Journal of Public Health* 99, 540-548.

Baake, M., Grimm, U., & Giegerich, R. (2006). Surprises in approximating Levenshtein distances. *Journal of Theoretical Biology* 243, 279-282.

Bainton, D., Moore, F., & Sweetnam, P. (1977). Temperature and deaths from ischaemic heart disease. *British Journal of Preventive and Social Medicine* 31, 49-53.

Bartley, M. & Blane, D. (1994) Appropriateness of deprivation indices must be ensured. *BMJ* 309, 1479.

Bartley, M., Sacker, A., Firth, D., & Fitzpatrick, R. (2000). Dimensions of inequality and the health of women; in *Understanding Health Inequalities*, H. Graham, ed. Buckingham, Open University Press, 58-74.

Bartley, M. (2004). *Health Inequality: an Introduction to Theories, Concepts and Methods*. Cambridge, Polity Press.

Beebe-Dimmer, J., Lynch, J., Turrell, G., Lustgarten, S., Raghunathan, T., & Kaplan, G. A. (2004). Childhood and adult socioeconomic conditions and 31-year mortality risk in women. *American Journal of Epidemiology* 159, 481-490.

Ben-Shlomo, Y. & Kuh, D. (2002). A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology* 31, 285-293.

Berney, L. R. & Blane, D. B. (1997). Collecting retrospective data: accuracy of recall after 50 years judged against historical records. *Social Science and Medicine* 45, 1519-1525.

Berry, H. L. (2008). Social capital elite, excluded participators, busy working parents and aging, participating less: types of community participators and their mental health. *Social Psychiatry and Psychiatric Epidemiology* 43, 527-537.

Bhopal, R. S. (2002). *Concepts of Epidemiology.* Oxford, Oxford University Press.

Blair-Loy, M. (1999). Career patterns of executive women in Finance: an optimal matching analysis. *American Journal of Sociology* 104, 1346-1397.

Blane, D. B. (1996). Collecting retrospective data: development of a reliable method and a pilot study of its use. *Social Science and Medicine* 42, 751-757.

Blane, D. B., Berney, L. R., Davey Smith, G., Gunnell, D. J., & Holland, P. (1999). Reconstructing the life course: health during early old age in a follow-up study based on the Boyd Orr cohort. *Public Health* 113, 117-124.

Blane, D., Mitchell, R., & Bartley, M. (2000). The "inverse housing law" and respiratory health. *Journal of Epidemiology & Community Health* 54, 745-749.

Blane, D. B. (2005). Cohort Profile: The Boyd Orr lifegrid sub-sample - medical sociology study of life course influences on early old age. *International Journal of Epidemiology* 34, 750-754.

Boardman, B. (2000). Introduction; in *Cutting the Cost of Cold: Affordable Warmth for Healthier Homes*, J. Rudge & F. Nicol, eds. London, E & FN Spon, 3-13.

Brennan, P. J., Greenberg, G., Miall, W. E., & Thompson, S. G. (1982). Seasonal variation in arterial blood pressure. *BMJ* 285, 919-923.

Brock, A. (2008). Excess winter mortality in England and Wales, 2007/08 (provisional) and 2006/07 (final); in *Health Statistics Quarterly 40 (Winter 2008)*, London, Office for National Statistics, 66-76.

Bronnum-Hansen, H. & Baadsgaard, M. (2007). Increasing social inequality in life expectancy in Denmark. *European Journal of Public Health* 17, 585-586.

Brown, J., Comber, M., Gibson, K., & Howard, S. (1985). Marriage and the family; in *Values and Social Change in Britain*, M. Abrams, D. Gerard, & N. Timms, eds. Basingstoke, The Macmillan Press Ltd., 109-145.

Brunekreef, B., Dockery, D. W., & Krzyzanowski, M. (1995). Epidemiologic studies on short-term effects of low levels of major ambient air pollution components. *Environmental Health Perspectives* 103 (Suppl2), 3-13.

Brunner, E., Shipley, M. J., Blane, D., Smith, G. D., & Marmot, M. (1999). When does cardiovascular risk start? Past and present socioeconomic circumstances and risk factors in adulthood. *Journal of Epidemiology & Community Health* 53, 757-764.

Bull, G. M. (1973). Meteorological correlates with myocardial and cerebral infarction and respiratory disease. *British Journal of Preventive and Social Medicine* 27, 108-113.

Ciccone, G., Forastiere, F., Agabiti, N., Biggeri, A., Bisanti, L., Chellini, E., Corbo, G., Dell'Orco, V., Dalmasso, P., Volante, T. F., Galassi, C., Piffer, S., Renzoni, E., Rusconi, F., Sestini, P., & Viegi, G. (1998). Road traffic and adverse respiratory effects in children. *Occupational and Environmental Medicine* 55, 771-778.

Clinch, J. P. & Healy, J. D. (2000). Housing standards and excess winter mortality. *Journal of Epidemiology & Community Health* 54, 719-720.

Coggins, J. M. (1999). Dissimilarity measures for clustering strings; in *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Reissue edn, D. Sankoff & J. Kruskal, eds. Stanford, CSLI Publications, 311-321.

Cole, T. (2007). The life course plot in life course analysis; in *Epidemiological Methods in Life Course Research*, A. Pickles, B. Maughan, & M. E. J. Wadsworth, eds. Oxford, Oxford University Press, 137-155.

Collins, M. (2000). Cold, cold housing and respiratory illness; in *Cutting the Cost of Cold: Affordable Warmth for Healthier Homes*, J. Rudge & F. Nicol, eds. London, E & FN Spon, 36-47.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society (Series A)* 134, 321-367.

Cotes, J. E., Chinn, D. J., & Miller, M. R. (2006). *Lung Function: Physiology, Measurement and Application in Medicine*, 6th edn, Oxford, Blackwell Publishing.

Crutchfield, J. P. & Young, K. (1989). Inferring statistical complexity. *Physical Review Letters* 63, 105-108.

Dalstra, J. A. A., Kunst, A. E., Mackenbach, J. P., & The EU Working Group on Socioeconomic Inequalities in Health (2006). A comparative appraisal of the relationship of education, income and housing tenure with less than good health among the elderly in Europe. *Social Science and Medicine* 62, 2046-2060.

Danet, S., Richard, F., Montaye, M., Beauchant, S., Lemaire, B., Graux, C., Cottel, D., Marecaux, N., & Amouyel, P. (1999). Unhealthy effects of atmospheric temperature and pressure on the occurrence of myocardial infarction and coronary deaths: a 10-year survey. *Circulation* 100, E1-E7.

Davey Smith, G., Bartley, M., & Blane, D. (1990). The Black report on socioeconomic inequalities in health 10 years on. *BMJ* 301, 373-377.

Davey Smith, G., Carroll, D., Rankin, S., & Rowan, D. (1992). Socioeconomic differentials in mortality: evidence from Glasgow graveyards. *BMJ* 305, 1554-1557.

Davey Smith, G., Blane, D. B., & Bartley, M. (1994). Explanations for socio-economic differentials in mortality: evidence from Britain and elsewhere. *European Journal of Public Health* 4, 131-144.

Davey Smith, G., Neaton, J. D., Wentworth, D., Stamler, R., & Stamler, J. (1996). Socioeconomic differentials in mortality risk among men screened for the Multiple Risk Factor Intervention Trial: I. white men. *American Journal of Public Health* 86, 486-496.

Davey Smith, G., Hart, C., Blane, D. B., Gillis, C., & Hawthorne, V. (1997). Lifetime socioeconomic position and mortality: prospective observational study. *BMJ* 314, 547-552.

Davey Smith, G., Hart, C., Watt, G., Hole, D., & Hawthorne, V. (1998). Individual social class, area-based deprivation, cardiovascular disease risk-factors and mortality: the Renfrew and Paisley study. *Journal of Epidemiology & Community Health* 52, 399-405.

Davey Smith, G., Dorling, D., & Shaw, M. (2001). *Poverty, Inequality and Health in Britain 1800-2000*. Bristol, The Policy Press.

Davey Smith, G. (2003). *Health Inequalities: Lifecourse Approaches*. Bristol, The Policy Press.

Department of Health and Social Security (1980). *Inequalities in Health: Report of a Research Working Group ('The Black Report')*. London, DHSS.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association* 263, 1385-1389.

Dockery, D. W. & Pope, C. A. (1994). Acute respiratory effects of particulate air pollution. *Annual Review of Public Health* 15, 107-132.

Donaldson, G. C., Tchernjavskii, V. E., Ermakov, S. P., Bucher, K., & Keatinge, W. R. (1998a). Winter mortality and cold stress in Yekaterinburg, Russia: interview survey. *BMJ* 316, 514-518.

Dong, G. & Pei, J. (2007). *Sequence Data Mining*. New York, Springer.

Dorling, D., Rigby, J., Wheeler, B., Ballas, D., Thomas, B., Fahmy, E., Gordon, D., & Lupton, R. (2007). *Poverty, Wealth and Place in Britain, 1968 to 2005*. Bristol, The Policy Press.

Dorward, A. J., Colloff, M. J., MacKay, N. S., McSharry, C., & Thomson, N. C. (1988). Effect of house dust mite avoidance measures on adult atopic asthma. *Thorax* 43, 98-102.

Drever, F. & Whitehead, M. (1997). *Health Inequalities: Decennial Supplement*. London, The Stationery Office.

Edelbrock, C. (1979). Mixture model tests of hierarchical clustering algorithms: the problem of classifying everybody. *Multivariate Behavioral Research* 14, 367-384.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1-26.

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1, 54-75.

Ellaway, A. & Macintyre, S. (1998). Does housing tenure predict health in the UK because it exposes people to different levels of housing related hazards in the home or its surroundings? *Health and Place* 4, 141-150.

Elzinga, C. H. (2009). Complexity of categorical time series. *Sociological Methods and Research*, in press (available at http://home.fsw.vu.nl/ch.elzinga/Complexity%20Preliminary.pdf [accessed 17/09/2009])

Erickson, B. W. & Sellers, P. H. (1999). Recognition of patterns in genetic sequences; in *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Reissue edn, D. Sankoff & J. Kruskal, eds. Stanford, CSLI Publications, 55-91.

Erikson, R. & Goldthorpe, J. H. (1992). *The Constant Flux*. Oxford, Clarendon Press.

Evans, J., Hyndman, S., Stewart-Brown, S., Smith, D., & Petersen, S. (2000). An epidemiological study of the relative importance of damp housing in relation to adult health. *Journal of Epidemiology & Community Health* 54, 677-686.

Evans, G. W. & Kantrowitz, E. (2002). Socioeconomic status and health: the potential role of environmental risk exposure. *Annual Review of Public Health* 23, 303-331.

Ferraty, F. & Vieu, P. (2006). *Nonparametric Functional Data Analysis*. New York, Springer.

Fewell, Z., Davey Smith, G., & Sterne, J. A. C. (2007). The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *American Journal of Epidemiology* 166, 646-655.

Fielding, A. H. (2006). *Cluster and Classification Techniques for the Biosciences*. Cambridge, Cambridge University Press.

Fogarty, M. (1985). British attitudes to work; in *Values and Social Change in Britain*, M. Abrams, D. Gerard, & N. Timms, eds. Basingstoke, The Macmillan Press Ltd., 173-200.

Ford, G., Ecob, R., Hunt, K., Macintyre, S., & West, P. (1994). Patterns of class inequality in health through the lifespan: class gradients at 15, 35 and 55 years in the West of Scotland. *Social Science and Medicine* 39, 1037-1050.

Freeman, G. H. & Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* 38, 141-149.

Fu, K.-S. & Lu, S.-Y. (1977). A clustering procedure for syntactic patterns. *IEEE Transactions on Systems, Man, and Cybernetics* 7, 734-742.

Fuller-Thomson, E., Hulchanski, J. D., & Hwang, S. (2000). The housing / health relationship: what do we know? *Reviews on Environmental Health* 15, 109-133.

Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2009). *Mining Sequence Data in R with the TraMineR Package: A User's Guide for Version 1.2*. Geneva, University of Geneva; URL: http://mephisto.unige.ch/pub/TraMineR/Doc/1.2/TraMineR-1.2-Users-Guide.pdf (accessed 10/08/2009)

Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J., & Davey Smith, G. (2006a). Indicators of socioeconomic position (part 1). *Journal of Epidemiology & Community Health* 60, 7-12.

Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J., & Davey Smith, G. (2006b). Indicators of socioeconomic position (part 2). *Journal of Epidemiology & Community Health* 60, 95-101.

Gemmell, I., McLoone, P., Boddy, F. A., Dickinson, G. J., & Watt, G. C. M. (2000). Seasonal variation in mortality in Scotland. *International Journal of Epidemiology* 29, 274-279.

Giaconi, S., Palombo, C., Genovesi-Ebert, A., Marabotti, C., Volterrani, D., & Ghioni, S. (1988). Long-term reproducibility and evaluation of seasonal influences on blood pressure monitoring. *Journal of Hypertension* 6 (suppl 4), S64-S66.

Gill, P. & de Wildt, G. (2003). *Housing and Health: The Role of Primary Care*. Abingdon, Radcliffe Medical Press.

Giussani, V. (1994). *The UK Clean Air Act 1956: An Empirical Investigation (CSERGE Working Paper GEC 94-20)*. London, Centre for Social and Economic Research on the Global Environment.

Glymour, M. M., Avendano, M., Haas, S., & Berkman, L. F. (2008). Lifecourse social conditions and racial disparities in incidence of first stroke. *Annals of Epidemiology* 18, 904-912.

Goldthorpe, J. H. (1980). *Social Mobility and Class Structure in Modern Britain*. Oxford, Clarendon Press.

Gordon, A. D. (1999). *Classification*, 2nd edn. Boca Raton (Florida), Chapman & Hall / CRC.

Gordon, D., Hyde, J., Trost, D. C., Whaley, F. S., Hannan, P. J., Jacobs, D. R., & Ekelund, L. (1988). Cyclic seasonal variation in plasma lipid and lipoprotein levels: the lipid research clinics coronary primary prevention trial placebo group. *Journal of Clinical Epidemiology* 41, 679-689.

Gotzsche, P. C. & Johansen, H. K. (2008). House dust mite control measures for asthma: systematic review. *Allergy* 63, 646-659.

Gower, J. C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* 3, 5-48.

Graham, H. (2002). Socio-economic change and inequalities in men and women's health in the UK; in *The Sociology of Health and Illness Reader*, S. Nettleton & U. Gustaffson, eds. Cambridge, Polity Press, 240-255.

Hallqvist, J., Lynch, J., Bartley, M., Lang, T., & Blane, D. (2004). Can we disentangle life course processes of accumulation, critical period and social mobility? An analysis of disadvantaged socio-economic positions and myocardial infarction in the Stockholm Heart Epidemiology Programme. *Social Science and Medicine* 58, 1555-1562.

Halpin, B. & Chan, T. W. (1998). Class careers as sequences: an optimal matching analysis of work-life histories. *European Sociological Review* 14, 111-130.

Harding, S., Bethune, A., & Rosato, M. (1997). Second study supports results of Whitehall study after retirement. *BMJ* 314, 1130.

Hardle, W. & Simar, L. (2007). *Applied Multivariate Statistical Analysis*, 2nd edn. Berlin, Springer-Verlag.

Hart, C., Davey Smith, G., & Blane, D. (1998). Inequalities in mortality by social class measured at 3 stages of the lifecourse. *American Journal of Public Health* 88, 471-474.

Hayes, B. C. & Miller, R. L. (1993). The silenced voice: female social mobility patterns with particular reference to the British Isles. *British Journal of Sociology* 44, 653-672.

Healy, J. D. (2003). Excess winter mortality in Europe: a cross country analysis identifying key risk factors. *Journal of Epidemiology & Community Health* 57, 784-789.

Hegewald, M. J. & Crapo, R. O. (2007). Socioeconomic status and lung function. *Chest* 132, 1608-1614.

von Hertzen, L. & Haahtela, T. (2009a). House dust mites in atopic diseases: accused for 45 years but not guilty? *American Journal of Respiratory and Critical Care Medicine* 180, 113-119.

von Hertzen, L. & Haahtela, T. (2009b). Rebuttal from Drs. von hertzen and Haahtela [letter]. *American Journal of Respiratory and Critical Care Medicine* 180, 120-121.

Heslop, P., Davey Smith, G., Macleod, J., & Hart, C. (2001). The socioeconomic position of employed women, risk factors and mortality. *Social Science and Medicine* 53, 477-485.

Hintze, J. L. & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *American Statistician* 52, 181-184.

Hopton, J. L., Platt, S. D., & Macleod, L. (2003). Housing conditions and health consequences; in *Housing and Health: the Role of Primary Care*, P. Gill & G. de Wildt, eds. Abingdon, Radcliffe Medical Press Ltd., 17-45.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edn. New York, John Wiley & Sons, Inc.

Howden-Chapman, P., Matheson, A., Crane, J., Viggers, H., Cunningham, M., Blakely, T., Cunningham, C., Woodward, A., Saville-Smith, K., O'Dea, D., Kennedy, M., Baker, M., Waipara, N., Chapman, R., & Davie, G. (2007). Effect of insulating existing houses on health inequality: cluster randomised study in the community. *BMJ* 334, 460-469.

Howieson, S. G. & Lawson, A. (2000). Dust mite allergens, indoor humidity and asthma; in *Cutting the Cost of Cold: Affordable Warmth for Healthier Homes*, J. Rudge & F. Nicol, eds. London, E & FN Spon, 62-77.

Howieson, S. G., Lawson, A., McSharry, C., Morris, G., McKenzie, E., & Jackson, J. (2003). Domestic ventilation rates, indoor humidity and dust mite allergens: are our homes causing the asthma pandemic? *Building Services Engineering Research and Technology* 24, 137-147.

Illsley, R. & Baker, D. (1991). Contextual variations in the meaning of health inequality. *Social Science and Medicine* 32, 359-365.

Institute for Environment and Health (IEH) (1996). *IEH Assessment on Indoor Air Quality in the Home: Nitrogen Dioxide, Formaldehyde, Volatile Organic Compounds, House Dust Mites, Fungi and Bacteria (Assessment A2).* Leicester, Institute for Environment and Health.

Institute for Environment and Health (IEH) (2001). *Indoor Air Quality in the Home: Final Report on DETR Contract EPG 1/5/12*. Leicester, Institute for Environment and Health.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys* 31, 264-323.

Johnson, H. & Griffiths, C. (2003). Estimating excess winter mortality in England and Wales. *Health Statistics Quarterly* 20, 19-24.

Johnson, R. A. & Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*, 3rd edn. Englewood Cliffs (New Jersey), Prentice-Hall Inc.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* 32, 241-254.

Jones, A. P. (1998). Asthma and domestic air quality. *Social Science and Medicine* 47, 755-764.

Kaplan, G. A. & Keil, J. E. (1993). Socioeconomic factors and cardiovascular disease: a review of the literature. *Circulation* 88, 1973-1998.

Katz, M. H. (2002). *Multivariable Analysis: a Practical Guide for Clinicians*, 1st corrected edn. Cambridge, Cambridge University Press.

Keatinge, W. R., Coleshaw, S. R. K., Cotter, F., Mattock, M., Murphy, M., & Chelliah, R. (1984). Increases in platelet and red cell counts, blood viscosity, and arterial pressure during mild surface cooling: factors in mortality from coronary and cerebral thrombosis in winter. *BMJ* 289, 1405-1408.

Keatinge, W. R. (1986). Seasonal mortality among people with unrestricted home heating. *BMJ* 293, 732-733.

Keatinge, W. R., Coleshaw, S. R. K., & Holmes, J. (1989). Changes in seasonal mortalities with improvement in home heating in England and Wales from 1964 to 1984. *International Journal of Biometeorology* 33, 71-76.

Keatinge, W. R. & Donaldson, G. C. (2000). Cold weather, cold homes and winter mortality; in *Cutting the Cost of Cold: Affordable Warmth for Healthier Homes*, J. Rudge & F. Nicol, eds. London, E & FN Spon, 17-24.

Kennedy, B. P., Kawachi, I., Glass, R., & Prothrow-Stith, D. (1998). Income distribution, socioeconomic status, and self rated health in the United States: multilevel analysis. *BMJ* 317, 917-921.

Khaw, K.-T. (1995). Temperature and cardiovascular mortality. *Lancet* 345, 337-338.

Koepsell, T. D. & Weiss, N. S. (2003). *Epidemiologic Methods: Studying the Occurrence of Illness*. Oxford, Oxford University Press.

Korsgaard, H. H. & Dahl, R. (1994). Clinical efficacy of reduction in house-dust mite exposure in specially designed, mechanically ventilated 'healthy' homes. *Allergy* 49, 866-870.

Krieger, N., Williams, D. R., & Moss, N. E. (1997). Measuring social class in US public health research: concepts, methodologies and guidelines. *Annual Review of Public Health* 18, 341-378.

Krieger, N., Chen, J. T., & Selby, J. V. (2001). Class inequalities in women's health: combined impact of childhood and adult social class - a study of 630 US women. *Public Health* 115, 175-185.

Krieger, J. & Higgins, D. L. (2002). Housing and health: time again for Public Health action. *American Journal of Public Health* 92, 758-768.

Kruskal, J. (1999). An overview of sequence comparison; in *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Reissue edn., D. Sankoff & J. Kruskal, eds. Stanford, CSLI Publications, 1-44.

Kruskal, J. & Liberman, M. (1999). The symmetric time-warping problem: from continuous to discrete; in *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Reissue edn, D. Sankoff & J. Kruskal, eds. Stanford, CSLI Publications, 125-161.

Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47, 583-621.

Kuehne, G., Bjornlund, H., & Cheers, B. (2008). Identifying common traits among Australian irrigators using cluster analysis. *Water Science and Technology* 58, 587-595.

Kuh, D. & Ben-Shlomo, Y. (1997). *A Life Course Approach to Chronic Disease Epidemiology*. Oxford, Oxford University Press.

Kuh, D., Ben-Shlomo, Y., Lynch, J., Hallqvist, J., & Power, C. (2003). Glossary: Life course epidemiology. *Journal of Epidemiology & Community Health* 57, 778-783.

Laaksonen, E., Martikainen, P., Lahelma, E., Lallukka, T., Rahkonen, O., Head, J., & Marmot, M. (2007). Socioeconomic circumstances and common mental disorders among Finnish and British public sector employees: evidence from the Helsinki Health Study and the Whitehall II Study. *International Journal of Epidemiology* 36, 776-786.

Land, K. C. & Nagin, D. S. (1996). Micro-models of criminal careers: a synthesis of the criminal careers and life course approaches via semiparametric mixed Poisson regression models, with empirical applications. *Journal of Quantitative Criminology* 12, 163-191.

Lavine, M. (2008). Mathematical techniques and the number of groups. *Behavioral and Brain Sciences* 31, 83-84.

Legendre, P. & Legendre, L. (1998). *Numerical Ecology*, 2nd edn. Amsterdam, Elsevier.

Li, M. & Vitanyi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd edn. New York, Springer.

Ljung, R. & Hallqvist, J. (2006). Accumulation of adverse socioeconomic position over the entire life course and the risk of myocardial infarction among men and women: results from the Stockholm Heart Epidemiology Program (SHEEP). *Journal of Epidemiology & Community Health* 60, 1080-1084.

Lloyd, S. (2001). Measures of complexity: a nonexhaustive list. *IEEE Control Systems Magazine* 21, 7-8.

Loucks, E. B., Lynch, J., Pilote, L., Fuhrer, R., Almeida, N. D., Hughes, R., Agha, G., Murabito, J. M., & Benjamin, E. J. (2009). Life-course socioeconomic position and incidence of coronary heart disease. *American Journal of Epidemiology* 169, 829-836.

Lynch, J. (1996). Social position and health. *Annals of Epidemiology* 6, 21-23.

Lynch, J., Kaplan, G. A., & Shema, S. J. (1997). Cumulative impact of sustained economic hardship on physical, cognitive, psychological, and social functioning. *New England Journal of Medicine* 337, 1889-1895.

Lynch, J., Davey Smith, G., Kaplan, G. A., & House, J. S. (2000). Income inequality and mortality: importance to health of individual income, psychosocial environment, or material conditions. *BMJ* 320, 1200-1204.

Lynch, J. & Davey Smith, G. (2005) A life course approach to chronic disease epidemiology. *Annual Review of Public Health* 26, 1-35.

MacIndoe, H. & Abbott, A. (2004). Sequence analysis and optimal matching techniques for social science data; in *Handbook of Data Analysis*, M. Hardy & A. Bryman, eds. London, Sage Publications Ltd., 387-406.

Macintyre, S. (1986). The patterning of health by social position in contemporary Britain: directions for sociological research. *Social Science and Medicine* 23, 393-415.

Macintyre, S. (1997). The Black Report and beyond: what are the issues? *Social Science and Medicine* 44, 723-745.

Macleod, J., Davey Smith, G., Metcalfe, C., & Hart, C. (2005). Is subjective social status a more important determinant of health than objective social status? Evidence from a prospective observational study of Scottish men. *Social Science and Medicine* 61, 1916-1929.

Manor, O., Matthews, S., & Power, C. (1997). Comparing measures of health inequality. *Social Science and Medicine* 45, 761-771.

Marchant, B., Ranjadayalan, K., Stevenson, R., Wilkinson, P., & Timmis, A. D. (1993). Circadian and seasonal factors in the pathogenesis of acute myocardial infarction: the influence of environmental temperature. *Heart* 69, 385-387.

Marmot, M., Ryff, C. D., Bumpass, L. L., Shipley, M., & Marks, N. F. (1997a). Social inequalities in health: next questions and converging evidence. *Social Science and Medicine* 44, 901-910.

Marmot, M., Bosma, H., Hemingway, H., Brunner, E., & Stansfeld, S. (1997b). Contribution of job control and other risk factors to social variations in coronary heart disease incidence. *Lancet* 350, 235-239.

Marmot, M. & Wilkinson, R. (2001). Psychosocial and material pathways in the relation between income and health: a response to Lynch et al. *BMJ* 322, 1233-1236.

Martin, C. J., Platt, S. D., & Hunt, S. M. (1987). Housing conditions and ill health. *BMJ* 294, 1125-1127.

Masters, J. (1961). *The Road Past Mandalay: a Personal Narrative*. London, Michael Joseph Ltd.

McCullagh, P. & Nelder, J. A. (1999). *Generalized linear models*, 2nd edn. Boca Raton (Florida), Chapman & Hall / CRC.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour; in *Frontiers in Econometrics*, P. Zarembka, ed. New York, Academic Press, 105-142.

McVicar, D. & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society (Series A)* 165, 317-334.

Meng, Y.-Y., Wilhelm, M., Rull, R. P., English, P., Nathan, S., & Ritz, B. (2008). Are frequent asthma symptoms among low-income individuals related to heavy traffic near homes, vulnerabilities, or both? *Annals of Epidemiology* 18, 343-350.

Mercer, J. B. (2003). Cold - an underrated risk factor for health. *Environmental Research* 92, 8-13.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika* 45, 325-342.

Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50, 159-179.

Mills, J. L. (1993). Data torturing. *New England Journal of Medicine* 329, 1196-1199.

Mishra, G., Nitsch, D., Black, S., De Stavola, B., Kuh, D., & Hardy, R. (2009). A structured approach to modelling the effects of binary exposure variables over the life course. *International Journal of Epidemiology* 38, 528-537.

Mitchell, M. (2009). *Complexity: a Guided Tour*. Oxford, Oxford University Press.

Mitchell, R., Blane, D., & Bartley, M. (2002). Elevated risk of high blood pressure: climate and the inverse housing law. *International Journal of Epidemiology* 31, 831-838.

Montnemery, P., Popovic, M., Andersson, M., Greiff, L., Nyberg, P., Lofdahl, C.-G., Svensson, C., & Persson, C. G. A. (2003). Influence of heavy traffic, city dwelling and socio-economic status on nasal symptoms assessed in a postal population survey. *Respiratory Medicine* 97, 970-977.

Murgatroyd, L. (1984). Women, men and the social grading of occupations. *British Journal of Sociology* 35, 473-497.

Naess, O., Claussen, B., Thelle, D. S., & Davey Smith, G. (2004a). Cumulative deprivation and cause specific mortality. A census based study of life course influences over three decades. *Journal of Epidemiology & Community Health* 58, 599-603.

Naess, O., Claussen, B., & Davey Smith, G. (2004b). Relative impact of childhood and adulthood socioeconomic conditions on cause specific mortality in men. *Journal of Epidemiology & Community Health* 58, 597-598.

Naess, O., Hernes, F. H., & Blane, D. B. (2006). Life-course influences on mortality at older ages: evidence from the Oslo Mortality Study. *Social Science and Medicine* 62, 329-336.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological Methods* 4, 139-157.

Nagin, D. S. & Tremblay, R. E. (1999). Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. *Child Development* 70, 1181-1196.

Nagin, D. S. (2005). *Group-based Modeling of Development*. Cambridge (Massachusetts), Harvard University Press.

National Academy of Sciences Institute of Medicine (NAS) (2000). *Clearing the Air: Asthma and Indoor Air Exposure*. Washington DC, National Academies Press.

National Institute of Standards and Technology (NIST) (2009a). Levenshtein distance (entry in *Dictionary of Algorithms and Data Structures*); URL: http://www.itl.nist.gov/div897/sqg/dads/HTML/Levenshtein.html (accessed 20/08/2009)

National Institute of Standards and Technology (NIST) (2009b). Kolmogorov complexity (entry in *Dictionary of Algorithms and Data Structures*); URL: http://www.itl.nist.gov/div897/sqg/dads/HTML/kolmogorov.html (accessed 20/08/2009)

Nyberg, F., Gustavsson, P., Jarup, L., Bellander, T., Berglind, N., Jakobsson, R., & Pershagen, G. (2000). Urban air pollution and lung cancer in Stockholm. *Epidemiology* 11, 487-495.

Office for National Statistics (2005). *The National Statistics Socio-economic Classification: User Manual*. Basingstoke, Palgrave Macmillan.

Office of the Deputy Prime Minister (2006). *Housing Health and Safety Rating System: Operating Guidance*. London, The Office of the Deputy Prime Minister.

Oreszczyn, T. & Pretlove, S. (2000). Mould Index; in *Cutting the Cost of Cold: Affordable Warmth for Healthier Homes*, J. Rudge & F. Nicol, eds. London, E & FN Spon, 122-133.

Pearce, N., Douwes, J., & Beasley, R. (2000). Is allergen exposure the major primary cause of asthma? *Thorax* 55, 424-431.

Peat, J. K., Dickerson, J., & Li, J. (1998). Effects of damp and mould in the home on respiratory health: a review of the literature. *Allergy* 53, 120-128.

Pensola, T. H. & Martikainen, P. (2003). Cumulative social class and mortality from various causes of adult men. *Journal of Epidemiology & Community Health* 57, 745-751.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ* 316, 1236-1238.

Pickles, A. & De Stavola, B. (2007). An Overview of Models and Methods for Life Course Analysis; in *Epidemiological Methods in Life Course Research*, A. Pickles, B. Maughan, & M. E. J. Wadsworth, eds. Oxford, Oxford University Press, 182-220.

Platts-Mills, T. A. E., de Weck, A. L., Aalberse, R. C., Bessot, J. C., Bjorksten, B., & Bischoff, E. (1989). Dust mite allergens and asthma - a worldwide problem. *Journal of Allergy and Clinical Immunology* 83, 416-427.

Platts-Mills, T. A. E., Erwin, E. A., Heymann, P. W., & Woodfolk, J.A. (2009). The evidence for a causal role of dust mites in asthma. *American Journal of Respiratory and Critical Care Medicine* 180, 109-113.

Power, C., Manor, O., & Matthews, S. (1999). The duration and timing of exposure: effects of socioeconomic environment on adult health. *American Journal of Public Health* 89, 1059-1065.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, 2nd edn. New York, Springer.

Ramsay, S. E., Whincup, P. H., Morris, R. W., Lennon, L. T., & Wannamethee, S. G. (2008). Extent of social inequalities in disability in the elderly: results from a population-based study of British men. *Annals of Epidemiology* 18, 896-903.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 846-850.

Reed, C. E. & Kita, H. (2004). The role of protease activation of inflammation in allergic respiratory disease. *Journal of Allergy and Clinical Immunology* 114, 997-1008.

Reuterwall, C., Hallqvist, J., Ahlbom, A., De Faire, U., Diderichsen, F., Hogstedt, C., Pershagen, G., Theorell, T., Wiman, B., Wolk, A., & The SHEEP Study Group (1999). Higher relative, but lower absolute risks of myocardial infarction in women than in men: analysis of some major risk factors in the SHEEP study. *Journal of Internal Medicine* 246, 161-174.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*, 2nd edn. Belmont (California), Duxbury Press.

Richardson, G., Eick, S., & Jones, R. (2005). How is the indoor environment related to asthma?: literature review. *Journal of Advanced Nursing* 52, 328-339.

Rogot, E., Sorlie, P. D., & Johnson, N. J. (1992). Life expectancy by employment status, income, and education in the National Longitudinal Mortality Study. *Public Health Reports* 107, 457-461.

Rohwer, G. & Potter, U. (2005). *TDA User's Manual*. Bochum, Ruhr-Universitat Bochum.

Rose, D. (1995). *Official Social Classifications in the UK (Social Research Update Issue 9)*. Guildford, University of Surrey.

Rose, D. & O'Reilly, K. (1998). *The ESRC Review of Government Social Classifications*. London, Office for National Statistics.

Rose, D., Pevalin, D. J., & O'Reilly, K. (2005). *The National Statistics Socio-economic Classification: Origins, Development and Use*. Basingstoke, Palgrave Macmillan.

Rothman, K. J. (1981). Induction and latent periods. *American Journal of Epidemiology* 114, 253-259.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology* 1, 43-46.

Sabbe, D., De Bourdeaudhuij, I., Legiest, E., & Maes, L. (2008). A cluster-analytical approach towards physical activity and eating habits among 10-year-old children. *Health Education Research* 23, 753-762.

Samet, J. M. (1987). Epidemiologic approaches for the identification of asthma. *Chest* 91, 74S-78S.

Sankoff, D. & Kruskal, J. (1999). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Reissue edn. Stanford, CSLI Publications.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York, Springer.

Savitz, D. A. & Olshan, A. F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology* 142, 904-908.

Schonnesson, L. N., Atkinson, A., Williams, M. L., Bowen, A., Ross, M. W., & Timpson, S. C. (2008). A cluster analysis of drug use and sexual HIV risks and their correlates in a sample of African-American crack cocaine smokers with HIV infection. *Drug and Alcohol Dependence* 97, 44-53.

Shannon, W. D. (2008). Cluster analysis; in *Handbook of Statistics - Volume 27,* C. R. Rao, J. P. Miller, & D. C. Rao, eds. Amsterdam, Elsevier, 342-366.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for Normality (complete samples). *Biometrika* 52, 591-611.

Shaw, M., Dorling, D., Gordon, D., & Davey Smith, G. (1999). *The widening gap: health inequalities and policy in Britain*. Bristol, The Policy Press.

Shaw, M. (2004). Housing and public health. *Annual Review of Public Health* 25, 397-418.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society (Series B)* 43, 310-313.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London, Chapman & Hall.

Singh-Manoux, A., Ferrie, J. E., Chandola, T., & Marmot, M. (2004). Socioeconomic trajectories across the life course and health outcomes in midlife: evidence for the accumulation hypothesis? *International Journal of Epidemiology* 33, 1072-1079.

Sneath, P. H. A. & Sokal, R. R. (1973). *Numerical Taxonomy: the Principles and Practice of Numerical Classification*. San Francisco, W.H. Freeman.

Sorensen, A. (1994). Women, family and class. *Annual Review of Sociology* 20, 27-47.

Spencer, F. A., Goldberg, R. J., Becker, R. C., & Gore, J. M. (1998). Seasonal distribution of acute myocardial infarction in the Second National Registry of Myocardial Infarction. *Journal of the American College of Cardiology* 31, 1226-1233.

Stevenson, J. (1990). *British Society 1914-45*. London, Penguin Books.

Stout, R. W. & Crawford, V. (1991). Seasonal variations in fibrinogen concentrations among elderly people. *Lancet* 338, 9-13.

Stovel, K., Savage, M., & Bearman, P. (1996). Ascription into achievement: Models of career systems at Lloyds Bank, 1890-1970. *American Journal of Sociology* 102, 385-399.

Stovel, K. (2001). Local sequential patterns: the structure of lynching in the Deep South, 1882-1930. *Social Forces* 79, 843-880.

Strachan, D. P. & Elton, R. A. (1986). Relationship between respiratory morbidity in children and the home environment. *Family Practice* 3, 137-142.

Strachan, D. P. (1988). Damp housing and childhood asthma: validation of reporting of symptoms. *BMJ* 297, 1223-1226.

Strachan, D. P. & Sanders, C. H. (1989). Damp housing and childhood asthma: respiratory effects of indoor air temperature and relative humidity. *Journal of Epidemiology & Community Health* 43, 7-14.

Strachan, D. P., Flannigan, B., McCabe, E. M., & McGarry, F. (1990). Quantification of airborne mould in the homes of children with and without wheeze. *Thorax* 45, 382-387.

Strachan, D. P. (1997). Respiratory and allergic diseases; in *A Life Course Approach to Chronic Disease Epidemiology*, D. Kuh & Y. Ben-Shlomo, eds. Oxford, Oxford University Press, 101-120.

Strachan, D. P. (2000). The role of environmental factors in asthma. *British Medical Bulletin*, vol. 56, no. 4, pp. 865-882.

Stronks, K. & Kunst, A. E. (2009). The complex interrelationship between ethnic and socio-economic inequalities in health. *Journal of Public Health* [advance access] doi:10.1093/pubmed/fdp070.

Stryer, D. B. & Browner, W. (1994). More on data torturing. *New England Journal of Medicine* 330, 861-862.

Sturgis, P. & Sullivan, L. (2008). Exploring social mobility with latent trajectory groups. *Journal of the Royal Statistical Society (Series A)* 171, 65-88.

Szklo, M. & Nieto, F. J. (2007). *Epidemiology: Beyond the Basics*, 2nd edn. Sudbury (Massachusetts), Jones & Bartlett.

Szreter, S. R. S. (1984). The genesis of the Registrar-General's social classification of occupations. *British Journal of Sociology* 35, 522-546.

The Eurowinter Group (1997). Cold exposure and winter mortality from ischaemic heart disease, cerebrovascular disease, and all causes in warm and cold regions of Europe. *Lancet* 349, 1341-1346.

Theodoridis, S. & Koutroumbas, K. (2006). *Pattern Recognition*, 3rd edn. London, Academic Press.

Thompson, J. R. (1998). Re: Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology* 147, 801-806.

Thomson, H., Petticrew, M., & Morrison, D. (2001). Health effects of housing improvement: systematic review of intervention studies. *BMJ* 323, 187-190.

Timm, N. H. (2002). *Applied Multivariate Analysis*. New York, Springer.

Townsend, P., Davidson, N., & Whitehead, M. (1992). *Inequalities in Health: The Black Report and The Health Divide*. Harmondsworth, Penguin Books.

Utell, M. J., Warren, J., & Sawyer, R. F. (1994). Public health risks from motor vehicle emissions. *Annual Review of Public Health* 15, 157-158.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th edn. New York, Springer.

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics* 27, 352-372.

Verhoeff, A. P. & Burge, H. A. (1997). Health risk assessment of fungi in home environments. *Annals of Allergy, Asthma and Immunology* 78, 544-556.

Voorhorst, R., Spieksma, F. Th. M., Varekamp, H., Leupen, M. J., & Lyklema, A. W. (1967). The house-dust mite (Dermatophagoides pteronyssinus) and the allergens it produces. Identity with the house-dust allergen. *Journal of Allergy* 39, 325-339.

Wadsworth, M. E. J. (1997). Health inequalities in the life course perspective *Social Science and Medicine* 44, 859-869.

Wadsworth, M. E. J., Maughan, B., & Pickles, A. (2007). Development and progression of life course ideas in epidemiology; in *Epidemiological Methods in Life Course Research*, A. Pickles, B. Maughan, & M. E. J. Wadsworth, eds. Oxford, Oxford University Press, 1-25.

Wagner, R. A. & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the Association for Computing Machinery* 21, 168-173.

Walker, J., Mitchell, R., Platt, S. D., Petticrew, M., & Hopton, J. L. (2006). Does usage of domestic heating influence internal environmental conditions and health? *European Journal of Public Health* 16, 463-469.

Walker, J., Mitchell, R., Platt, S., Blane, D. (2008). Representing trajectories of social location over the lifecourse: a challenge and a proposed solution [conference abstract]. *Journal of Epidemiology and Community Health* 62, A2.

Walker, J., Mitchell, R., Petticrew, M., & Platt, S. D. (2009). The effects on health of a publicly funded domestic heating programme: a prospective controlled study. *Journal of Epidemiology & Community Health* 63, 12-17.

Wallace, R. B. (2007). Epidemiology and public health; in *Public Health and Preventive Medicine*, 15th revised edn. R. B. Wallace, N. Kohatsu, & J. M. Last, eds. London, McGraw-Hill, 5-26.

Walshaw, M. J. & Evans, C. C. (1986). Allergen avoidance in house dust mite sensitive adult asthma. *Quarterly Journal of Medicine* 58, 199-215.

Wannamethee, S. G. & Shaper, G. (1997). Socioeconomic status within social class and mortality: a prospective study in middle-aged British men. *International Journal of Epidemiology* 26, 532-541.

Watt, H. C., Carson, C., Lawlor, D. A., Patel, R., & Ebrahim, S. (2009). Influence of life course socioeconomic position on older women's health behaviors: findings from the British Women's Heart and Health Study. *American Journal of Public Health* 99, 320-327.

Webb, A. R. (2002). *Statistical Pattern Recognition*, 2nd edn. Chichester, John Wiley & Sons Ltd.

West, R. (1989). Seasonal variation in CHD mortality. *International Journal of Epidemiology* 18, 463-464.

Whitehead, M. (1992). *The Health Divide*. Harmondsworth, Penguin Books.

Whyte, K. F., & Flenley, D. C. (1986). House dust sensitivity and asthma. *Quarterly Journal of Medicine* 226, 89-93.

Wilhelmsen, L., Svardsudd, K., Korsan-Bengtsen, K., Larsson, B., Welin, L., & Tibblin, G. (1984). Fibrinogen as a risk factor for stroke and myocardial infarction. *New England Journal of Medicine* 311, 501-505.

Wilkinson, P., Landon, M., & Stevenson, S. (2000). Housing and winter death: epidemiological evidence; in *Cutting the Cost of Cold: Affordable Warmth for Healthier Homes*, J. Rudge & F. Nicol, eds. London, E & FN Spon, 25-35.

Williams, D. R. (1990). Socioeconomic differentials in health: a review and redirection. *Social Psychology Quarterly* 53, 81-99.

Williams, D. R. & Collins, C. (1995). US socioeconomic and racial differences in health: patterns and explanations. *Annual Review of Sociology* 21, 349-386.

Woodhouse, P. R., Khaw, K.-T., & Plummer, M. (1993). Seasonal variation of blood pressure and its relationship to ambient temperature in an elderly population. *Journal of Hypertension* 11, 1267-1274.

Yarnell, J. W., Baker, I. A., Sweetnam, P. M., Bainton, D., O'Brien, J. R., Whitehead, P. J., & Elwood, P. C. (1991). Fibrinogen, viscosity, and white blood cell count are major risk factors for ischemic heart disease. The Caerphilly and Speedwell collaborative heart disease studies. *Circulation* 83, 836-844.

Yoder, J. A., Glenn, B. D., Benoit, J. B., & Zettler, L. W. (2007). The giant Madagascar hissing-cockroach (*Gromphadorhina portentosa*) as a source of antagonistic moulds: concerns arising from its use in a public setting. *Mycoses* 51, 95-98.